Meta's Broken Promises

Systemic Censorship of Palestine Content on Instagram and Facebook

Meta's Broken Promises

Systemic Censorship of Palestine Content
on Instagram and Facebook

Human Rights Watch defends the rights of people worldwide. We scrupulously investigate abuses, expose the facts widely, and pressure those with power to respect rights and secure justice. Human Rights Watch is an independent, international organization that works as part of a vibrant movement to uphold human dignity and advance the cause of human rights for all.

Human Rights Watch is an international organization with staff in more than 40 countries, and offices in Amsterdam, Beirut, Berlin, Brussels, Chicago, Geneva, Goma, Johannesburg, London, Los Angeles, Nairobi, New York, Paris, San Francisco, Sydney, Tokyo, Toronto, Tunis, Washington DC, and Zurich.

For more information, please visit our website: http://www.hrw.org

Meta's Broken Promises

Systemic Censorship of Palestine Content on Instagram and Facebook

Summary

Meta's policies and practices have been silencing voices in support of Palestine and Palestinian human rights on Instagram and Facebook in a wave of heightened censorship of social media amid the hostilities between Israeli forces and Palestinian armed groups that began on October 7, 2023. This systemic online censorship has risen against the backdrop of unprecedented violence, including an estimated 1,200 people killed in Israel, largely in the Hamas-led attack on October 7, and over 18,000 Palestinians killed as of December 14, largely as a result of intense Israeli bombardment.

Between October and November 2023, Human Rights Watch documented over 1,050 takedowns and other suppression of content Instagram and Facebook that had been posted by Palestinians and their supporters, including about human rights abuses. Human Rights Watch publicly solicited cases of any type of online censorship and of any type of viewpoints related to Israel and Palestine. Of the 1,050 cases reviewed for this report, 1,049 involved peaceful content in support of Palestine that was censored or otherwise unduly suppressed, while one case involved removal of content in support of Israel. The documented cases include content originating from over 60 countries around the world, primarily in English, all of peaceful support of Palestine, expressed in diverse ways. This distribution of cases does not necessarily reflect the overall distribution of censorship. Hundreds of people continued to report censorship after Human Rights Watch completed its analysis for this report, meaning that the total number of cases Human Rights Watch received greatly exceeded 1,050.

Human Rights Watch found that the censorship of content related to Palestine on Instagram and Facebook is systemic and global. Meta's inconsistent enforcement of its own policies led to the erroneous removal of content about Palestine. While this appears to be the biggest wave of suppression of content about Palestine to date, Meta, the parent company of Facebook and Instagram, has a well-documented record of overbroad crackdowns on content related to Palestine. For years, Meta has apologized for such overreach and promised to address it. In this context, Human Rights Watch found Meta's behavior fails to meet its human rights due diligence responsibilities. Despite the censorship documented in this report, Meta allows a significant amount of pro-Palestinian expression and denunciations of Israeli government policies. This does not, however,

excuse its undue restrictions on peaceful content in support of Palestine and Palestinians, which is contrary to the universal rights to freedom of expression and access to information.

This report builds on and complements years of research, documentation, and advocacy by Palestinian, regional, and international human rights and digital rights organizations, in particular 7amleh, the Arab Center for the Advancement of Social Media, and Access Now.

In reviewing the evidence and context associated with each reported case, Human Rights Watch identified six key patterns of undue censorship, each recurring at least 100 times, including 1) removal of posts, stories, and comments; 2) suspension or permanent disabling of accounts; 3) restrictions on the ability to engage with content—such as liking, commenting, sharing, and reposting on stories—for a specific period, ranging from 24 hours to three months; 4) restrictions on the ability to follow or tag other accounts; 5) restrictions on the use of certain features, such as Instagram/Facebook Live, monetization, and recommendation of accounts to non-followers; and 6) "shadow banning," the significant decrease in the visibility of an individual's posts, stories, or account, without notification, due to a reduction in the distribution or reach of content or disabling of searches for accounts.

In addition, dozens of users reported being unable to repost, like, or comment on Human Rights Watch's post calling for evidence of online censorship, which was flagged as "spam." Instagram accounts posted about the call for censorship documentation in comments with an email address to send Human Rights Watch evidence. Instagram removed the comments, citing a violation of its Community Guidelines.

Human Rights Watch's analysis of the cases suggests four underlying, systemic factors that contributed to the censorship:
1) Flaws in Meta policies, principally its Dangerous Organizations and Individuals (DOI) policy, which bars organizations or individuals "that proclaim a violent mission or are engaged in violence" from its platforms. Understandably, the policy prohibits incitement to violence. However, it also contains sweeping bans on vague categories of speech, such as "praise" and "support" of "dangerous organizations," which it relies heavily on the United States government's

designated lists of terrorist organizations to define. The US list includes political movements that have armed wings, such as Hamas and the Popular Front for the Liberation of Palestine. The ways in which Meta enforces this policy effectively bans many posts that endorse major Palestinian political movements and quells the discussion around Israel and Palestine;

2) Inconsistent and opaque application of Meta policies, in particular on exceptions for newsworthy content, that is, content that Meta allows to remain visible in the public interest, even if it is otherwise violates their policies;

3) Apparent deference to requests by governments for content removals, such as requests by Israel's Cyber Unit and other countries' internet referral units to remove content; and

4) Heavy reliance on automated tools for content removal to moderate or translate Palestine-related content.

In addition, in over 300 cases documented by Human Rights Watch, users reported and provided evidence of being unable to appeal the restriction on their account to the platform, which left the user unable to report possible platform violations and without any access to an effective remedy.

Meta has long been on notice that its policies have resulted in the silencing of Palestinian voices and their supporters on its platforms. The evidence of censorship documented in this report stems from the same concerns that human and digital rights organizations raised on previous occasions, such as in 2021, when the planned takeovers by Israeli authorities of Palestinian homes in the Sheikh Jarrah neighborhood of occupied East Jerusalem triggered protests and violence along with censorship of pro-Palestine content on Facebook and Instagram. In a 2021 report, Human Rights Watch documented Facebook's censorship of the discussion of rights issues pertaining to Israel and Palestine and warned that Meta was "silencing many people arbitrarily and without explanation, replicating online some of the same power imbalances and rights abuses that we see on the ground."

In response to years of digital and human rights organizations calling for an independent review of Meta's content moderation policies and a 2021 recommendation from Meta's Oversight Board—an external body created by Meta to appeal content moderation

decisions and to provide non-binding policy guidance—Meta commissioned Business for Social Responsibility (BSR), an independent entity, to investigate whether Facebook had applied its content moderation in Arabic and Hebrew, including its use of automation, without bias. In September 2022, BSR published "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," which found that Meta's actions "appear to have had an adverse human rights impact…on the rights of Palestinian users to freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred."

Based on recommendations from the Oversight Board, the BSR report, and engagement with civil society over the years, Meta made several commitments to addressing concerns around Palestine-related censorship. However, Meta's practices during the hostilities that erupted in October 2023 show that the company has not delivered on the promises it made two years ago. As this report demonstrates, the problem has grown only more acute.

Under the United Nations Guiding Principles on Business and Human Rights (UNGPs), companies have a responsibility to avoid infringing on human rights, identify and address the human rights impacts of their operations, and provide meaningful access to a remedy to those whose rights they abused. For social media companies, including Meta, this responsibility includes aligning their content moderation policies and practices with international human rights standards, ensuring that decisions to take down content are transparent and not overly broad or biased, and enforcing their policies consistently.

Meta should permit protected expression, including about human rights abuses and political movements, on its platforms. It should begin by overhauling its Dangerous Organizations and Individuals policy so that it comports with international human rights standards. Meta should also audit its enforcement of its "newsworthy allowances," to ensure that these are being applied in an effective, equitable, and non-discriminatory manner.

The company should improve transparency around requests by governments' internet referral units, including Israel's Cyber Unit, to remove content "voluntarily"—that is, without a court or administrative order to do so—and about its use of automation and machine learning algorithms to moderate or translate Palestine-related content. It should

carry out due diligence on the human rights impact of temporary changes to its recommendation algorithms that it introduced in response to the hostilities between Israel and Hamas since October 7. Meta should also take urgent steps to work with civil society to set targets for the implementation of its outstanding commitments to address overreach in its content suppression of Palestine-related content.

Methodology

Human Rights Watch conducted the research for this report between October and November 2023. In October, Human Rights Watch published a call for evidence of online censorship—used here and in other Human Rights Watch reporting in its colloquial sense of improper limitations on or suppression of free expression—and suppression of content related to Israel and Palestine on social media since October 7, which we posted in English, Arabic, and Hebrew from the main Human Rights Watch accounts on Instagram, X (formerly known as Twitter), and TikTok.[1] Human Rights Watch attempted to solicit information from its entire global audience.

Human Rights Watch requested the following information be sent via email, to an address dedicated to this research, from social media users who reported experiencing censorship: screenshots of the original content, the relevant social media platform, the date and country from which the content was posted, the form of censorship experienced (removal, "shadow ban", disabling features, inability to engage with content, etc.), the notification from the relevant platform (if any), prior engagement figures (in case of shadow banning), the account URL, appeal status (if any), and any other relevant information. In addition to the cases we received via solicitation, people spontaneously sent us cases and we identified several additional publicly available cases for inclusion.

Human Rights Watch solicited cases of any type of online censorship and of any type of viewpoint related to Israel and Palestine. Of the 1,050 cases reviewed for this report, 1,049 cases documented involved examples of online censorship and suppression of content in support of Palestine, while one case contained an example of removal of content in support of Israel.[2] This distribution of cases does not necessarily reflect the overall distribution of censorship.

---

[1] Human Rights Watch posted the call for evidence in English and Arabic on Instagram, X, and TikTok, and posted the call for evidence in Hebrew on X.

[2] The one case in support of Israel was on Instagram, of a post that included the statement, "No ceasefire until all hostages are home. No ceasefire until Hamas is destroyed. Israel has a right to defend itself. Palestinians have a right to be free from Hamas oppression."

Human Rights Watch's internal data shows that the call for evidence posted on social media reached audiences across the globe, including in Israel. Most of the content that Human Rights Watch received was in English[3] and originated from the following countries and territories: Antigua and Barbuda, Australia, Austria, Bahrain, Bangladesh, Belgium, Bolivia, Bosnia, Brazil, Brunei, Canada, Congo, Croatia, Denmark, Egypt, Finland, France, Germany, Ghana, India, Indonesia, Ireland, Israel, Italy, Jordan, Kenya, Kuwait, Lebanon, Libya, Lithuania, Malaysia, Mexico, Netherlands, New Zealand, Norway, Oman, Pakistan, Palestine, Panama, Peru, Portugal, Puerto Rico, Qatar, Romania, Singapore, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Thailand, Trinidad, Tunisia, Türkiye, the United Kingdom, and the United States.

The researchers reviewed all 1,285 reports of online censorship received via email by November 28, 2023, either in response to our solicitation or spontaneously submitted. We excluded cases in which there was insufficient evidence to substantiate the claim of censorship or that or that did not include content about Israel or Palestine. We also screened evidence for any speech that could be considered incitement to violence, discrimination, or hostility by evaluating the content of the post, the context around the post (other comments, media, etc.) and additional information provided by the person who reported the censorship, and notifications from Meta. The researchers used a combination of evidence provided by the user including screenshots and background material in the email and publicly available information to assess whether the claim of unjustified restrictions on their content or account by Meta was substantiated. If the researchers did not have enough information to fully assess the context of the post to confirm that the content was peaceful support for Palestine or Palestinians, we excluded the case.

This analysis identified a data set of 1,050 cases of censorship, i.e. the removal or suppression of protected expression. This dataset understates the volume of censorship that was reported to us, as hundreds of people continued to report instances of censorship after our November 28 cutoff. At time of writing, we have received a total number of 1,736 reports. While these additional cases are not included in this report's analysis, a review of these reports indicates hundreds of more instances wherein support of Palestine or

---

[3] In addition to English, Human Rights Watch received 27 cases of censored content in Arabic, Danish, Dutch, French, German, and Swedish.

Palestinians was censored. This distribution of cases does not necessarily reflect the overall distribution of censorship.

Most reports received and evidence documented by Human Rights Watch were about postings on Instagram and Facebook (with fewer overall instances being reported about X, TikTok, and other platforms). Meta platforms (Instagram and Facebook) have had high usage rates, both in response to hostilities in Israel and Palestine since October 7 as well as historically. Facebook and Instagram each have the highest usage rates, with over 3 billion and over 2.3 billion monthly active users respectively, compared to other platforms such as X (close to 400 million), Telegram (over 800 million), and TikTok (over 1 billion), as of 2023.[4]

This report is an analysis of the verified cases of content removal we received or documented and is not a global comparative analysis of overall censorship of political statements and viewpoints. The trends identified in these cases are not intended to reflect the general distribution of censorship across social media platforms. The findings of this report, namely that Meta's censorship primarily suppressed protected expression in support of Palestine or Palestinians on Instagram and Facebook, pertain to trends only within these 1,050 cases.

The researchers anonymized all the information social media users shared with Human Rights Watch, and assured the people who reported their experiences that none of their information would be shared or published without their explicit and informed consent. None of the participants in the research received any compensation.

---

[4] Meta, "Meta Reports Third Quarter 2023 Results," October 25, 2023, https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Third-Quarter-2023-Results/default.aspx (accessed December 12, 2023); Sherry Fairchok, "Bank of America Leads Among US Banks in Instagram Engagement," *Insider Intelligence*, August 14, 2023, https://www.insiderintelligence.com/content/bank-of-america-leads-among-us-banks-instagram-engagement (accessed December 12, 2023), Kali Hays, "X Downloads Hit Lowest Level in Over a Decade as Use Nears a Yearly Low Under Elon Musk's Ownership," *Business Insider*, September 9, 2023, https://www.businessinsider.com/twitter-downloads-usage-sinks-under-elon-musk-ownership-2023-9 (accessed December 12, 2023); "Thanks a Billion!," TikTok Newsroom, September 27, 2021, https://newsroom.tiktok.com/en-africa/a-billion-users-on-tiktok (accessed December 12, 2023); "TikTok showcases how its Thriving Business Community is Driving Creativity and Impact at Cannes Lions Festival," TikTok Newsroom, June 19, 2023, https://newsroom.tiktok.com/en-ca/tiktok-showcases-how-its-thriving-business-community-is-driving-creativity-and-impact-at-cannes-lions-festival-ca (accessed December 12, 2023).

Human Rights Watch wrote to Meta on November 15, 2023, to share the findings of our research and to solicit Meta's perspective. Meta's response is reflected in this report. Human Rights Watch also wrote to the Israeli Cyber Unit at Israel's Office of the State Attorney on November 22, to request information on the unit's requests to social media companies since October 7 and justification for making such requests. At time of writing, the Cyber Unit had not responded. All letters and responses are included in full in this report's annex.

The research included consultations with several human rights and digital rights organizations including 7amleh, Access Now, and Amnesty International.

I. Background

Political Context: The Broader Environment for Censorship

Palestinians today are facing unprecedented levels of violence and repression. The Israeli military's current operations in Gaza began following an unprecedented Hamas-led attack on Israel on October 7, 2023, in which an estimated 1,200 people were killed and more than 200 people were taken hostage, according to Israeli authorities.[5][6] As of December 14, 2023, around 18,700 Palestinians had been killed in Gaza, including around 7,700 children, according to authorities in Gaza.[7] Israel cut off essential services to Gaza and prevented the entry of all but a trickle of aid.[8]

The extreme violence and dire humanitarian situation have made it harder to seek and impart information. According to preliminary investigations by the Committee to Protect Journalists (CPJ), as of December 17, 2023, at least 64 journalists and media workers were confirmed dead: 57 Palestinian, 4 Israeli, and 3 Lebanese.[9] CPJ said that the first month of the hostilities in Israel and Gaza marked "the deadliest month for journalists" since they began documenting journalist fatalities in 1992.[10] Additionally, repeated and prolonged communications blackouts in Gaza have impeded access to reliable and lifesaving information, as well as the ability to document and share evidence of human rights abuses.[11]

---

[5] "Hamas, Islamic Jihad: Holding Hostages is a War Crime," Human Rights Watch news release, October 19, 2023, https://www.hrw.org/news/2023/10/19/hamas-islamic-jihad-holding-hostages-war-crime.

[6] "Israel/Palestine: Videos of Hamas-Led Attacks Verified," Human Rights Watch news release, October 18, 2023, https://www.hrw.org/news/2023/10/18/israel/palestine-videos-hamas-led-attacks-verified.

[7] United Nations Office for the Coordination of Humanitarian Affairs (OCHA), "Hostilities in the Gaza Strip and Israel | Flash Update #70," December 15, 2023, https://www.unocha.org/publications/report/occupied-palestinian-territory/hostilities-gaza-strip-and-israel-flash-update-70 (accessed December 17, 2023).

[8] "Israel: Immediately Restore Electricity, Water, Aid to Gaza," Human Rights Watch news release, October 21, 2023, https://www.hrw.org/news/2023/10/21/israel-immediately-restore-electricity-water-aid-gaza.

[9] Committee to Protect Journalists (CPJ), "Journalist casualties in the Israel-Gaza war," December 17, 2023, https://cpj.org/2023/12/journalist-casualties-in-the-israel-gaza-conflict/ (accessed December 17, 2023).

[10] Committee to Protect Journalists (CPJ), "Israel-Gaza War," undated, https://cpj.org/full-coverage-israel-gaza-war/ (accessed December 18, 2023).

[11] "Gaza: Communications Blackout Imminent Due to Fuel Shortage," Human Rights Watch news release, November 15, 2023, https://www.hrw.org/news/2023/11/15/gaza-communications-blackout-imminent-due-fuel-shortage.

Against this backdrop, the broader environment for free expression about Palestine is under increasing pressure. While the focus of this report is censorship of social media content, online censorship does not exist in a vacuum. On November 23, 2023, United Nations experts issued a statement expressing alarm at a worldwide wave of attacks, reprisals, criminalization, and sanctions against those who publicly express solidarity with the victims of the hostilities between Israeli forces and Palestinian armed groups.[12] The experts noted that artists, academics, journalists, activists, and athletes have faced particularly harsh consequences and reprisals from states and private actors because of their prominent roles and visibility.[13] Protecting free expression on issues related to Israel and Palestine is especially important considering the shrinking space for discussion.[14]

On November 8, 2023 the Israeli Knesset passed an amendment to the country's Counter-Terrorism Law of 2016 that makes the "consumption of terrorist materials" a criminal offense. Adalah, an Israeli human rights organization, criticized the law as "[o]ne of the most intrusive and draconian legislative measures ever passed by the Israeli Knesset," saying it "invades the realm of personal thoughts and beliefs and significantly amplifies state surveillance of social media use."[15] Even before the amendment's passage, it was reported on October 17, 2023 that Israel's police stated at least 170 Palestinians had been arrested or brought in for questioning since the Hamas attack on the basis of online expression.[16] Adalah has documented 251 instances between October 7, 2023 and November 13, 2023 where people had been issued a warning, interrogated, or, in at least 132 cases, arrested for activity that [as it classifies] largely falls within the right to freedom of expression. The organization described this as "widespread and coordinated" efforts to repress expression of dissent against the Israeli government's attack on Gaza.[17] The Israeli

---

[12] Office of the United Nations High Commissioner for Human Rights, "Speaking out on Gaza / Israel must be allowed: UN experts," November 23, 2023, https://www.ohchr.org/en/press-releases/2023/11/speaking-out-gaza-israel-must-be-allowed-un-experts (accessed November 30, 2023).

[13] Ibid.

[14] Sophia Goodfriend, "Israel's 'thought police' law ramps up dangers for Palestinian social media users," *972 Magazine*, November 24, 2023, https://www.972mag.com/israel-thought-police-surveillance-palestinians/(accessed November 30, 2023).

[15] Adalah, "Israeli Knesset Passes Draconian Amendment to the Counter-Terrorism Law Criminalizing "Consumption of Terrorist Publications," November 8, 2023, https://www.adalah.org/en/content/view/10951 (accessed November 30, 2023).

[16] Ghousoon Bisharat, Oren Ziv and Baker Zoubi, "'This is political persecution': Israel cracks down on internal critics of its Gaza war," *972 Magazine*, October 17, 2023, https://www.972mag.com/israel-gaza-war-political-persecution/(accessed December 6, 2023).

[17] Adalah, "Interrogations, Arrests and Indictments of Palestinian citizens of Israel over the last month", November 13, 2023, https://www.adalah.org/uploads/uploads/Criminal_Proceedings_Report_Eng_Nov_13.pdf (accessed December 12, 2023).

government's systematic oppression of millions of Palestinians, coupled with inhumane acts committed as part of a policy to maintain the domination by Jewish Israelis over Palestinians, amount to the crimes against humanity of apartheid and persecution.[18]

Palestinian authorities in the West Bank and Gaza have also clamped down on free expression,[19] while in several other countries, including the United States and across Europe, governments and private entities have taken steps to restrict the space for some forms of advocacy in support of Palestine.

Since October 7, 2023, artists, cultural workers, and academics in various countries have faced significant consequences in the form of silencing, censorship, and intimidation by some governments and private institutions as a result of non-violent, pro-Palestinian speech. [20] These include undue pressure or restrictions on academic freedom,[21] and Palestinian experts being disinvited from media interviews and conferences.[22] There have also been restrictions on peaceful protests in support of Palestine.[23] The punishing tactics against those expressing solidarity with Palestinians or criticizing Israeli war crimes in Gaza pose serious challenges to freedom of expression in a time of crisis and polarization over events on and since October 7.

Palestine Legal, an organization that protects the civil and constitutional rights of people in the United States who speak out for Palestinian freedom, received over 600 requests for

[18] Human Rights Watch, *A Threshold Crossed Israeli Authorities and the Crimes of Apartheid and Persecution*, (New York: Human Rights Watch, 2021), https://www.hrw.org/report/2021/04/27/threshold-crossed/israeli-authorities-and-crimes-apartheid-and-persecution.

[19] Human Rights Watch, *Two Authorities, One Way, Zero Dissent*, (New York: Human Rights Watch, 2018), https://www.hrw.org/report/2018/10/23/two-authorities-one-way-zero-dissent/arbitrary-arrest-and-torture-under.

[20] @AJEnglish, tweet with video, *X*, November 18, 2023, https://x.com/AJEnglish/status/1725997309031247890?s=20 (accessed November 30, 2023).

[21] "US rights group urges colleges to protect free speech amid Israel-Gaza war," *Al Jazeera*, November 1, 2023, https://www.aljazeera.com/news/2023/11/1/us-rights-group-urges-colleges-to-protect-free-speech-amid-gaza-war (accessed November 30, 2023).

[22] Chris McGreal, "Pro-Palestinian views face suppression in US amid Israel-Hamas war," *The Guardian*, October 21, 2023, https://www.theguardian.com/us-news/2023/oct/21/israel-hamas-conflict-palestinian-voices-censored (accessed November 30, 2023).

[23] "Israel-Palestine Hostilities Affect Rights in Europe," Human Rights Watch news release, October 26, 2023, https://www.hrw.org/news/2023/10/26/israel-palestine-hostilities-affect-rights-europe.

support between October 7, 2023 and November 15, 2023.[24] The organization reported "seeing an unprecedented wave of workplace discrimination," having received over 280 allegations involving employment concerns, over 60 of which involved people who said they had already been terminated from their jobs.[25]

In addition, authorities in various European countries have at times imposed excessive restrictions on pro-Palestine protest and speech since October 7, 2023.[26] French authorities placed a blanket ban on pro-Palestinian protests, a move overturned by the Council of State, France's highest administrative court, on October 18, 2023.[27] Before the decision, French authorities had banned 64 protests about Palestine, media reported.[28]

Since October, authorities in Germany have banned some pro-Palestinian protests while allowing others to take place,[29] prompting concern from the country's antisemitism commissioner, who noted that "demonstrating is a basic right."[30] On October 13, 2023, education authorities in Berlin gave schools permission to ban students from wearing the Palestinian *keffiyeh* (checkered black and white scarf) and displaying "free Palestine" stickers, raising concerns about the right to free expression and possible discrimination.[31]

---

[24] @pal_legal, "THREAD: Since Oct 7, Palestine Legal has received 600+ requests for support from advocates for Palestinian rights. The firings, campus repression, government calls for investigation—this is the greatest threat to free expression & political dissent since the McCarthy era." *X*, November 15, 2023, https://x.com/pal_legal/status/1724899909839864298?s=20 (accessed November 30, 2023).

[25] Ibid.

[26] "Israel-Palestine Hostilities Affect Rights in Europe," Human Rights Watch news release, October 26, 2023, https://www.hrw.org/news/2023/10/26/israel-palestine-hostilities-affect-rights-europe.

[27] "Top court rules France cannot ban pro-Palestinian rallies outright," *Radio France Internationale*, October 18, 2023, https://www.rfi.fr/en/france/20231018-top-court-to-rule-whether-france-s-ban-on-pro-palestinian-rallies-is-legal (accessed November 30, 2023).

[28] "Manifestation pro-Palestine à Paris : le tribunal administratif lève l'interdiction de la prefecture," *Ouest-France*, October 19, 2023, https://www.ouest-france.fr/monde/palestine/manifestation-pro-palestine-a-paris-le-tribunal-administratif-leve-linterdiction-de-la-prefecture-96b81ba4-6ea1-11ee-97d6-d90367762e60 (accessed November 30, 2023).

[29] Erika Solomon, "Germany's Stifling of Pro-Palestinian Voices Pits Historical Guilt Against Free Speech," *The New York Times*, November 10, 2023, https://www.nytimes.com/2023/11/10/world/europe/germany-pro-palestinian-protests.html (accessed November 30, 2023).

[30] Ashifa Kassam, "Rise in antisemitism 'brings Germans back to most horrific times'," *The Guardian*, October 24, 2023, https://www.theguardian.com/world/2023/oct/24/rise-in-antisemitism-brings-germans-back-to-most-horrific-times (accessed November 30, 2023).

[31] Senatsverwaltung für Bildung, Jugend und Familie, "Umgang mit Störungen des Schulfriedens im Zusammenhang mit dem Terrorangriff auf Israel" [Dealing with disruptions to school peace in connection with the terrorist attack on Israel], October 13, 2023, p. 2, https://mediendienst-integration.de/fileadmin/Dateien/Informationsschreiben_Umgang_mit_Sto__rungen_des_Schulfriedens.pdf (accessed November 30, 2023).

Bans on pro-Palestinian protests have been reported in Austria,[32] Hungary,[33] and Switzerland.[34]

In the United Kingdom, police in London have generally taken a nuanced approach to pro-Palestinian protest since October 7, including in relation to the use of slogans in protests that have been cited elsewhere in Europe to justify bans.[35] This is despite political pressure on the police by the then- UK home secretary, who called for use of "the full force of the law" in the context of pro-Palestinian protests [36] and a statement by the previous UK foreign minister, who in October called on pro-Palestinian supporters to "stay at home."[37] The then-UK immigration minister stated that visitors to the country will be "removed" if they "incite antisemitism," even if their conduct falls "below the criminal standard."[38]

Meta's Broken Promises

Meta has long been aware that its policies have resulted in the silencing of Palestinian voices and their supporters on its platforms. For years, digital rights and human rights organizations from the region, in particular 7amleh, have been documenting and calling Meta's attention to the disproportionately negative impact of its content moderation on

---

[32] "'From the river to the sea' prompts Vienna to ban pro-Palestinian protest," *Reuters*, October 11, 2023, https://www.reuters.com/world/from-river-sea-prompts-vienna-ban-pro-palestinian-protest-2023-10-11/ (accessed November 30, 2023).

[33] Zách Dániel, "They wanted to protest for the Palestinians, not Hamas, but at Orbán's word, the police banned the demonstration immediately," *Telex*, October 18, 2023, https://telex.hu/english/2023/10/18/they-wanted-to-protest-in-support-of-the-palestinians-not-hamas-but-at-orbans-word-the-police-immediately-banned-the-demonstrations (accessed November 30, 2023).

[34] "Hundreds turn out for unauthorised pro-Palestine rally in Zurich," *Swiss Info*, October 21, 2023, https://www.swissinfo.ch/eng/politics/unauthorised-pro-palestine-rally-marches-in-zurich/48911646 (accessed November 30, 2023).

[35] "Met response to terror attacks in Israel and ongoing military action in Gaza," *Metropolitan Police News*, October 20, 2023, https://news.met.police.uk/news/update-met-response-to-terror-attacks-in-israel-and-ongoing-military-action-in-gaza-474080 (accessed November 30, 2023).

[36] Elena Salvoni, "Suella Braverman urges police chiefs to use 'full force of the law' against shows of support for Hamas and warns waving Palestinian flag on British streets 'may not be legitimate'," *Daily Mail*, October 10, 2023, https://www.dailymail.co.uk/news/article-12615887/Suella-Braverman-urges-police-chiefs-force-law-against-shows-support-Hamas-warns-waving-Palestinian-flag-British-streets-not-legitimate.html (accessed November 30, 2023).

[37] "Israel-Hamas war: Foreign Secretary James Cleverly calls on pro-Palestinian protesters to stay at home," *Sky News*, October 10, 2023, https://news.sky.com/story/israel-hamas-war-foreign-secretary-james-cleverly-calls-on-pro-palestinian-protesters-to-stay-at-home-12981481 (accessed November 30, 2023).

[38] Aletha Adu, "Visitors to UK who incite antisemitism will be removed, says minister," *The Guardian*, October 25, 2023, https://www.theguardian.com/uk-news/2023/oct/25/visitors-to-uk-who-incite-antisemitism-will-be-removed-says-minister-robert-jenrick (accessed November 30, 2023).

Palestinians.[39] Human Rights Watch and others have pointed the company to underlying structural problems, such as flaws in its Dangerous Organizations and Individuals (DOI) policy,[40] inconsistent and opaque enforcement of its rules, influence by various governments over voluntary content removals, and heavy reliance on automation for content removal, moderation, and suppression.[41] When violence escalates in the region and people turn to social media to document, discuss, raise awareness around and condemn human rights abuses, and engage in political debate, the volume of content skyrockets, as do the levels and severity of censorship.

The events of May 2021 are emblematic of this dynamic. When plans by Israeli authorities to take over Palestinian homes in the Sheikh Jarrah neighborhood of occupied East Jerusalem triggered protests and escalation in violence in parts of Israel and the Occupied Palestinian Territories (OPT),[42] people experienced heavy handed censorship when they used social media to speak out. On May 7, 2021, a group of 30 human rights and digital rights organizations denounced social media companies for "systematically silencing users protesting and documenting the evictions of Palestinian families from their homes in the neighborhood of Sheikh Jarrah in Jerusalem."[43] In October 2021, Human Rights Watch published a report that documented Facebook's censorship of the discussion of rights issues pertaining to Israel and Palestine and warned that Meta was "silencing many people arbitrarily and without explanation, replicating online some of the same power imbalances and rights abuses that we see on the ground."[44]

---

[39] 7amleh - The Arab Center for the Advancement of Social Media, "Facebook and Palestinians: Biased or Neutral Content Moderation Policies?," October 29, 2018, https://7amleh.org/2018/10/29/7amleh-releases-policy-paper-facebook-and-palestinians-biased-or-neutral-content-moderation-policies (accessed November 30, 2023).

[40] Meta, "Facebook Community Standards: Dangerous Organizations and Individuals," https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed November 30, 2023).

[41] "Israel/Palestine: Facebook Censors Discussion of Rights Issues," Human Rights Watch news release, October 8, 2021, https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues; Electronic Frontier Foundation, "Tell Facebook: Stop Silencing Palestine," undated, https://stopsilencingpalestine.com (accessed December 1, 2023).

[42] Omar Shakir, "Jerusalem to Gaza, Israeli Authorities Reassert Domination," commentary, Human Rights Watch witness piece, May 11, 2021, https://www.hrw.org/news/2021/05/11/jerusalem-gaza-israeli-authorities-reassert-domination.

[43] Access Now, "Sheikh Jarrah: Facebook and Twitter systematically silencing protests, deleting evidence," May 7, 2021, https://www.accessnow.org/press-release/sheikh-jarrah-facebook-and-twitter-systematically-silencing-protests-deleting-evidence/ (accessed November 30, 2023).

[44] Israel/Palestine: Facebook Censors Discussion of Rights Issues," Human Rights Watch news release, October 8, 2021, https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues.

At the time, Facebook acknowledged several issues affecting Palestinians and their content, as well as those speaking about Palestinian matters globally,[45] some of which it attributed to "technical glitches"[46] and human error.[47] However, these issues did not explain the range of restrictions and suppression of content that Human Rights Watch observed. In a letter to Human Rights Watch, Facebook said it had already apologized for "the impact these actions have had on [Meta's] community in Palestine and on those speaking about Palestinian matters globally."[48]

In response to years of calls by digital and human rights organizations[49] and a recommendation from Meta's Oversight Board[50] for an independent review of the company's content moderation policies, Meta commissioned an independent investigation to determine whether Facebook's content moderation in Arabic and Hebrew, including its use of automation, had been applied without bias. In September 2022, Business for Social Responsibility (BSR), a "sustainable business network and consultancy", published its findings in its report "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021."[51] Among other findings, the report concluded that Meta's actions "appear to have had an adverse human rights impact on the rights of Palestinian users to freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred." The BSR report identified various instances where Meta policy and practice, combined with broader external dynamics, led

[45] Letter from Neil Potts VP, Trust and Safety Policy Facebook to Human Rights Watch, July 27, 2021, https://www.hrw.org/sites/default/files/media_2021/11/Letter%20AG%20HRW%20IsrPal%20%28003%29_Redacted.pdf

[46] Elizabeth Dwoskin, Gerrit De Vynck, "Facebook's AI treats Palestinian activists like it treats American Black activists. It blocks them." *Washington Post,* May 29, 2021, https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/ (accessed November 30, 2023).

[47] Ryan Mac, "Instagram Censored Posts About One Of Islam's Holiest Mosques, Drawing Employee Ire," *Buzzfeed News*, May 12, 2021, https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque (accessed November 30, 2023).

[48] Letter from Neil Potts VP, Trust and Safety Policy Facebook to Human Rights Watch, July 27, 2021, https://www.hrw.org/sites/default/files/media_2021/11/Letter%20AG%20HRW%20IsrPal%20%28003%29_Redacted.pdf

[49] Electronic Frontier Foundation, "Tell Facebook: Stop Silencing Palestine," undated, https://stopsilencingpalestine.com (accessed December 1, 2023).

[50] Meta, "Oversight Board Selects a case related to an Al Jazeera post on tensions between Israel and Palestine," June 12, 2023, https://transparency.fb.com/oversight/oversight-board-cases/al-jazeera-post-tensions-israel-palestine/ (accessed December 16, 2023).

[51] Business for Social Responsibility (BSR), "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (accessed November 30, 2023).

to different human rights impacts on Palestinian users as well as other Arabic-speaking users.[52]

In response to recommendations from the Oversight Board, the BSR report, and engagement with civil society over the years, Meta made several commitments to addressing concerns around Palestine-related censorship. However, the evidence of censorship and suppression of content about, and in support of, Palestinians and their supporters that Human Rights Watch documents in this report stems from the same underlying concerns that surfaced in 2021 and earlier, and shows that Meta has not delivered on the promises it made two years ago.

For example, Meta committed to improving its DOI policy, its rule on content on terrorist and other "dangerous entities," which is at the center of many of the content takedowns and account restrictions concerning Palestine. The DOI policy prohibits "organizations or individuals that proclaim a violent mission or are engaged in violence [from having] a presence on Meta."[53]

The DOI policy also bans "praise" and "substantive support" of groups or individuals from Meta's platforms. These are vague and broad terms that can include expression that is protected under international human rights law. In its scope and application, the DOI policy effectively bans the endorsement of many major Palestinian political movements and quells the discussion of current hostilities.[54] Meta had, in 2022, agreed to amend the policy to allow people to engage in social and political discourse more freely, consider removing its prohibition on "praise" of DOI entities, and make the penalties incurred under the DOI policy proportional to the violation. Meta says that it has completed its rollout of the social and political discourse carveout[55] and aims to launch changes to its "praise"

---

[52] Ibid.

[53] Meta, "Facebook Community Standards: Dangerous Organizations and Individuals," https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed November 30, 2023).

[54] See section "A "Dangerous" Policy for Public Debate" for further analysis of the DOI policy.

[55] Meta, "Meta Q3 2023 Quarterly Update on the Oversight Board," November 2023, https://transparency.fb.com/sr/meta-quarterly-update-q3-2023 (accessed December 9, 2023).

and its penalty system in the first half of 2024,[56] which suggests that "praise" will continue to be part of the policy.

The cases Human Rights Watch documents in this report indicate that Meta's erroneous removal of pro-Palestinian views has resulted in censorship of social and political discourse and content documenting or reacting to hostilities in Israel and the occupied Palestinian territories (OPT), including Gaza, sharing on-the-ground developments, or expressing solidarity with Palestinians.

Meta committed in August[57] and October 2021[58] to increasing transparency around government requests for content removals under its Community Standards, such as those from Israel's Cyber Unit, as well as internet referral units (IRUs) in other countries. IRU requests are prone to abuse because they risk circumventing legal procedures, lack transparency and accountability, and fail to provide users with access to an effective remedy. According to media reports on November 14, Israel's Cyber Unit sent Meta and other platforms 9,500 content takedown requests since October 7, 2023, 60 percent of which went to Meta.[59] Platforms are reported to have responded with a 94-percent compliance rate, according to an Israeli official. Two years after its commitment to increasing transparency, Meta has made no meaningful progress in informing its users or other members of the public how government requests are influencing what content is removed from Instagram and Facebook.

Meta also committed, in 2021, to provide greater transparency to users around its enforcement actions, including limiting certain features and reducing the visibility of accounts during user online searches, and communicating enforcement actions clearly.

---

[56] Meta, "Meta Update: Israel and Palestine Human Rights Due Diligence," September 2023, https://humanrights.fb.com/wp-content/uploads/2023/09/September-2023-Israel-and-Palestine-HRDD-Meta-Update.pdf (accessed December 9, 2023).

[57] Meta, "Case regarding the support of Abdullah Öcalan, founder of the PKK," June 12, 2023, https://transparency.fb.com/oversight/oversight-board-cases/support-of-abdullah-ocalan-founder-of-the-pkk/ (accessed November 30, 2023).

[58] Meta, "Oversight Board Selects a case related to an Al Jazeera post on tensions between Israel and Palestine," June 12, 2023, https://transparency.fb.com/en-gb/oversight/oversight-board-cases/al-jazeera-post-tensions-israel-palestine (accessed November 30, 2023).

[59] Thomas Brewster, "Israel Has Asked Meta And TikTok To Remove 8,000 Posts Related To Hamas War," *Forbes,* November 14, 2023, https://www.forbes.com/sites/thomasbrewster/2023/11/13/meta-and-tiktok-told-to-remove-8000-pro-hamas-posts-by-israel/?sh=16af3f16f6ce (accessed November 30, 2023).

Yet, a largely recurrent complaint Human Rights Watch received in researching this report was that users lost account features without warning.

As this report demonstrates, Meta's broken promises have led it to not only replicate past patterns of abuses, but also to amplify them. Censoring the voices and narratives of Palestinians and those voicing solidarity with them does not just impact those whose posts and accounts are restricted. It reduces the information to which the rest of the world has access regarding developments in Israel and Palestine at a time when the United Nations Secretary-General and UN human rights experts are warning with increasing urgency that Palestinians in Gaza are facing a humanitarian catastrophe.[60] Meta's failure to take decisive action in response to recommendations of its own Oversight Board, notwithstanding years of engagement with civil society, means that the company has failed to meet its human rights responsibilities.

---

[60] See, for example, "UN Secretary General Invokes Article 99 on Gaza," *Al Jazeera,* December 7, 2023, https://www.aljazeera.com/news/2023/12/7/un-secretary-general-invokes-article-99-on-gaza (accessed December 14, 2023); United Nations Office of the High Commissioner for Human Rights, "Gaza: UN experts call on international community to prevent genocide against the Palestinian people," November 16, 2023, https://www.ohchr.org/en/press-releases/2023/11/gaza-un-experts-call-international-community-prevent-genocide-against (accessed November 30, 2023).

II. Main Findings

Since October 7, Human Rights Watch has documented over 1,000 cases of unjustified takedowns and other suppression of content on Instagram and Facebook related to Palestine and Palestinians, including about human rights abuses. These cases detail various forms of censorship of posts and accounts documenting, condemning, and raising awareness about the unprecedented and ongoing hostilities in Israel and Palestine. The censorship of content related to Palestine on Instagram and Facebook is systemic, global, and a product of the company's failure to meet its human rights due diligence responsibilities.

The documented cases include content originating from over 60 countries around the world, primarily in English,[61] which carried a diversity of messages while sharing a singular characteristic: peaceful expression in support of Palestine or Palestinians.

In reviewing the evidence and context associated with each reported case, Human Rights Watch identified key patterns of censorship, each recurring in at least a hundred instances, including: 1) removal of posts, stories, and comments; 2) suspension or permanent disabling of accounts; 3) restrictions on the ability to engage with content—such as liking, commenting, sharing, and reposting on stories —for a specific period, ranging from 24 hours to three months; 4) restrictions on the ability to follow or tag other accounts; 5) restrictions on the use of certain features, such as Instagram/Facebook Live, monetization, and the recommendation of accounts to non-followers; and 6) "shadow banning," defined as the significant decrease in the visibility of an individual's posts, stories, or account, without notification, due to a reduction in the distribution or reach of content or disabling of searches for accounts.

Some users reported multiple forms of restrictions occurring simultaneously on their account. For example, some users who had their comments removed for violating Meta's spam policy—which prohibits content that is designed to deceive, or that attempts to

---

[61] In addition to English, the cases received by Human Rights Watch included cases of censored content in Arabic, Danish, Dutch, French, German, and Swedish.

mislead users, to increase viewership[62]—and were then unable to comment on any posts. In some cases, these users also reported their suspicion of being "shadow banned," based on perceived lower views of and engagement with their content. Some users provided evidence that Meta failed to specify which of their policies was violated.[63]

Throughout the research period for this report, Human Rights Watch received cases on a rolling basis, and the same users sometimes reported subsequent platform restrictions, indicating a gradual escalation in the type of restriction imposed on their content or account. For example, repeated comment removals were followed by restrictions in accessing features such as "live streaming," and a warning that the account could be suspended or permanently disabled. The more "strikes"[64] the user collected, the faster the next restriction on their content or account would become. One user described the pattern:

> I noticed a lot of my comments on Instagram were automatically removed as being "spam." At first the process of being marked as spam seemed to happen a few hours after I made the comments, and the next day it was nearly instantaneous. Then I could no longer "like" news posts about Palestine—I would try more than a dozen times and it would never work. I could "like" other stories posted by this same user. Eventually, I could not even respond to comments made on my own posts.[65]

In addition, most people who reported cases to Human Rights Watch said it was their first time experiencing restrictions on Meta's platforms since they joined years earlier. In every case, the censorship was strictly related to pro-Palestinian content since October 7. Some

---

[62] Meta, "Facebook Community Standards: Spam," https://transparency.fb.com/en-gb/policies/community-standards/spam/ (accessed December 6, 2023).

[63] For example, Evidence ID 64439, comment removed by Instagram on October 27 for violating unspecified community guideline; Evidence ID 1OEvo, comments removed by Instagram on October 31 for violating unspecified community guideline; Evidence ID 1A1oR, comment removed by Instagram on October 25 for violating unspecified community guideline; Evidence ID WGoK5, comments removed by Instagram on October 25 and October 30 for violating unspecified community guidelines; Evidence ID 4S6oN, comments removed by Instagram on October 20, October 24, October 30, November 4, and November 7, each for violating unspecified community guidelines; Evidence ID M2211, indicating deletion and preventing posting of hyperlinks in posts by Instagram; Evidence ID 3JeJ1, comments removed by Instagram on October 25 and twice on November 6 for violations of unspecified community guidelines; Evidence ID oNGX1, stories removed by Instagram on November 11 and November 12 for violating unspecified community guidelines.

[64] Meta, "Counting strikes," October 4, 2022, https://transparency.fb.com/enforcement/taking-action/counting-strikes (accessed December 16, 2023); Meta, "Restricting accounts," February 23, 2023, https://transparency.fb.com/enforcement/taking-action/restricting-accounts/ (accessed December 16, 2023).

[65] Evidence ID: H9BA2, Activity occurred on Instagram, October 30, 2023.

users reported examples of abusive content that incited violence or constituted hate speech against Palestinians remaining online while seemingly peaceful content advocating for Palestinian human rights was removed, at times on the same post. For example, to express outrage about abusive comments she experienced on Instagram, a user posted an Instagram "story"[66]—with a screenshot of a message addressed to her that said, "I wish Hamas will catch you, rape you slowly for hours and then kill you, while sending a video of this to your parents, just like they did to us," as well as her response, "If I knocked your glasses off right now you wouldn't even be able to see." The story was flagged and removed under Instagram's Guidelines on "violence or dangerous organizations."[67]

Over time, users who reported cases to Human Rights Watch said this led them to change their online behavior or engagement to adapt to and circumvent restrictions, effectively self-censoring to avoid accruing penalties imposed by the platform. Users described this as contributing to resentment at what they perceived as injustice or bias by the company. One person said they did not appeal the takedown to Meta because, "I do not want to put myself [on] their [Meta's] radar."[68] Instagram users also employ coded language, such as deliberate misspellings and symbols, in part to try to evade platform censorship resulting from automated moderation of content related to Palestine.

In many instances, users said they did not receive a warning or notification that their account was suspended or disabled or that Meta had barred their use of certain features. In cases of suspected "shadow banning," users said they were never informed by the platform that their content visibility was diminished. While some claims of shadow banning were supported with compelling evidence that their account was "shadow banned,"[69] many users concluded that they had been "shadow banned" based on a "hunch" or after noticing sudden changes in the number of views on their stories.

---

[66] "Stories" are a feature on Facebook and Instagram that lets users create and share photos and videos that will disappear after 24 hours. "Stories", Instagram Help Center, undated, https://help.instagram.com/1660923094227526/ (accessed December 12, 2023).

[67] Evidence ID: L45OD, Activity occurred on Instagram, October 17, 2023. Appeal filed with Oversight Board October 22, 2023.

[68] Evidence ID: CP39O, Activity occurred on Instagram, October 20, 2023.

[69] Evidence ID: W49Z6, Activity occurred on Instagram, November 2, 2023.

On October 18, 2023, Meta said that it fixed a "bug" that had significantly reduced reach on Stories that re-shared Reels and Feed posts on Instagram.[70] Yet, users continued to report and document shadow banning cases after that date. Due to Meta's lack of transparency around the issue of shadow banning, the parameters of the restriction remain unclear, and because users are not informed of any action taken on their account or content, the user is left without a remedy.[71]

In cases where removal or restrictions on content and accounts were accompanied by a notice to the user, Meta's most widely cited reasons were Community Guidelines (Instagram) or Standards (Facebook) violations, specifically those relating to "Dangerous Organizations and Individuals (DOI),[72] "adult nudity and sexual activity," "violent and graphic content," and "spam."[73] Among those violations, the most recurring policy invoked by Instagram and Facebook in the cases documented by Human Rights Watch was the "spam" policy. In reviewing these cases, Human Rights Watch found repeated instances of likely erroneous application of the "spam" policy that resulted in the censorship of Palestine-related content.

Human Rights Watch also found repeated inaccurate application of the "adult nudity and sexual activity" policy for content related to Palestine. In every one of the cases, we reviewed where this policy was invoked, the content included images of dead Palestinians over ruins in Gaza that were clothed, not naked. For example, multiple users reported their Instagram stories being removed under this policy when they posted the same image of a

---

[70] Meta, "Meta's Ongoing Efforts Regarding the Israel-Hamas War," October 13, 2023 (Updated: December 7, 2023), https://about.fb.com/news/2023/10/metas-efforts-regarding-israel-hamas-war/ (accessed November 30, 2023).

[71] According to the BSR report, violations of Meta policies like incitement to violence, hate speech, or bullying and harassment result in restrictions, such as reduced searchability of the account (i.e., requiring users to enter the exact name of the users account in order to find it rather than a normal keyword search) or reduced content visibility (i.e., placing content lower in feeds). Meta notifies users when their searchability has been impacted, but not when content visibility is reduced. Business for Social Responsibility, "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (accessed November 30, 2023), FN 11, page 5.

[72] For more information on violations related to "dangerous organizations and individuals," see below section.

[73] Meta, "Facebook Community Standards: Adult Nudity and Sexual Activity", https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/ (accessed December 12, 2023) ; Meta, "Facebook Community Standards: Violent and Graphic Content", https://transparency.fb.com/policies/community-standards/violent-graphic-content/ (accessed December 12, 2023); Meta, "Facebook Community Standards: Spam," https://transparency.fb.com/en-gb/policies/community-standards/spam/ (accessed December 6, 2023). Meta's "Violent and Graphic Content" policy includes an exception for "discussions about important issues such as human rights abuses, armed conflicts or acts of terrorism." In these cases, it allows graphic content (with some limitations) to help people to condemn and raise awareness about these situations.

Palestinian father in Gaza who was killed while he was holding his clothed daughter, who was also killed.

While "hate speech," "bullying and harassment," and "violence and incitement" policies[74] were less commonly invoked in the cases Human Rights Watch documented, the handful of cases where they were applied stood out as erroneous. For example, a Facebook user post that said, "How can anyone justify supporting the killing of babies and innocent civilians…" was removed under Community Standards on "bullying and harassment."[75] Another user posted an image on Instagram of a dead child in a hospital in Gaza with the comment, "Israel bombs the Baptist Hospital in Gaza City killing over 500…" which was removed under Community Guidelines on "violence and incitement."[76]

In over 300 cases documented by Human Rights Watch, the user reported and provided evidence of being unable to appeal the restriction on their account to the platform (Instagram or Facebook), indicating that the "Tell Us" button either did not work or did not lead anywhere when clicked, and the "Think that we've made a mistake?" option was disabled or unavailable. This left the user unable to report possible platform violations and without any access to an effective remedy.

---

[74] Meta, "Facebook Community Standards: Hate Speech", https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed December 12, 2023); Meta, "Facebook Community Standards: Bullying and Harassment", https://transparency.fb.com/policies/community-standards/bullying-harassment/ (accessed December 12, 2023); Meta, "Facebook Community Standards: Violence and Incitement", https://transparency.fb.com/policies/community-standards/violence-incitement/ (accessed December 12, 2023).

[75] Evidence ID RN3K0, comment posted to Facebook "How can anyone justify supporting the killing of babies and innocent civilians. I wouldn't wish it on any children, babies or innocent people" on October 17, 2023, and was removed for violating Facebook's Community Standards on bullying and harassment on October 20, 2023.

[76] Evidence ID M8LZ3, comment posted to Instagram on October 17, 2023. Full comment in the post was: "Israel bombs the Baptist Hospital in Gaza City killing over 500. A place where displaced Palestinians have sought refuge where the wounded are being treated…killing doctors and nurses administering lifesaving treatments. War crimes after war crime." Attribution of the Baptist/ al-Ahli Hospital explosion was a topic of fierce public debate. Human Rights Watch assessed that the explosion that killed and injured many civilians at al-Ahli Arab Hospital in Gaza on October 17, 2023, resulted from an apparent rocket-propelled munition, such as those commonly used by Palestinian armed groups, that hit the hospital grounds. Evidence available to Human Rights Watch makes the possibility of a large air-dropped bomb, such as those Israel has used extensively in Gaza, highly unlikely. See "Gaza: Findings on October 17 al-Ahli Hospital Explosion: Evidence Points to Misfired Rocket but Full Investigation Needed", Human Rights Watch news release, October 17, 2023, https://www.hrw.org/news/2023/11/26/gaza-findings-october-17-al-ahli-hospital-explosion.

Illustrative Examples

"From the River to the Sea"

The slogan "From the river to the sea, Palestine will be free" has reverberated at protests in solidarity with Palestinians around the world. In hundreds of cases documented by Human Rights Watch, this slogan, as well as comments such as "Free Palestine," "Ceasefire Now," and "Stop the Genocide," were repeatedly removed by Instagram and Facebook under "spam" Community Guidelines or Standards without appearing to take into account the context of these comments. These statements and the context in which they are used are clearly not spam nor appear to violate any other Facebook or Instagram Community Guidelines or Standards. For instance, the words in each of these statements on their face do not constitute incitement to violence, discrimination, or hostility. Meta has not offered a specific explanation as to why the context in which these statements appear would justify removal. In dozens of cases, the content removal was accompanied by platform restrictions on users' ability to engage with any other content on Instagram and Facebook, at times for prolonged periods.

Palestinian Flag Emoji

The Palestinian flag symbol, used frequently around the world to express solidarity with Palestine, has been subject to censorship on Instagram and Facebook. In one case, an Instagram user received a warning that the comment she posted "may be hurtful to others." The comment, which Human Rights Watch reviewed, consisted of nothing more than a series of Palestinian flag emojis.[77] In other cases, Meta hid the Palestinian flag from comment sections or removed it on the basis that it "harasses, targets, or shames others."[78] In October, Instagram apologized for adding "terrorist" to some Palestinian user public profiles who used the Arabic word "alhamdulillah" ("praise be to God") and the Palestinian flag emoji. Meta said the censorship was caused by a bug.[79] The issue arose when Instagram's auto translation feature translated bios that had the word "Palestinian"

---

[77] Evidence ID BIGL7, Activity occurred on Instagram, October 22, 2023.

[78] Sam Biddle, "Instagram Hid a Comment. It was Just Three Palestinian Flag Emojis," The Intercept, October 28, 2023, https://theintercept.com/2023/10/28/instagram-palestinian-flag-emoji/ (accessed November 30, 2023).

[79] Samantha Cole, "Instagram Is Hiding Its 'Palestinian Terrorists' Translation Problem Inside a Black Box," *404 Media*, https://www.404media.co/instagram-palestinian-arabic-translation-terrorists-ai/ (accessed November 30, 2023).

in Arabic, the Palestinian flag emoji, and the word "alhamdulillah" alongside one another as "Palestinian terrorists."[80]

Meta spokesperson Andy Stone confirmed to the US online media outlet *The Intercept* that the company has been hiding comments that contain the Palestinian flag emoji in certain "offensive" contexts that violate the company's rules. He added that Meta has not created any new policies specific to flag emojis. Asked about the contexts in which Meta hides the Palestinian flag, Stone pointed to the DOI policy, which designates Hamas as a terrorist organization, and cited a section of the Community Standards rulebook that prohibits any content "praising, celebrating or mocking anyone's death." The Palestinian flag pre-dates the existence of Hamas, which has its own distinct flag. Stone stated that Meta does not have a different standard to enforce rules with respect to the Palestinian flag emoji.[81]

Mention of "Hamas" Censored

Human Rights Watch documented hundreds of cases where the mere neutral mention of Hamas on Instagram and Facebook triggered the DOI policy, prompting the platforms to immediately remove posts, stories, comments, and videos, and restrict accounts that posted them. While the DOI policy[82] permits reporting on, neutrally discussing, or condemning designated organizations or individuals, it also states that "[if] a user's intention is ambiguous or unclear, we default to removing content." All cases that Human Rights Watch reviewed found that Meta removed even neutral mentions of Hamas in relation to developments in Gaza.

Suspension and Removal of Prominent Palestinian Accounts

Instagram and Facebook have in several instances since October 7 suspended or permanently disabled the accounts of prominent Palestinian content creators, independent Palestinian journalists, and Palestinian activists. Palestinian journalist Ahmed Shihab-Eldin reported on November 18, 2023 that he lost access five times to his Instagram account, which has nearly one million followers, since October 7. Shihab-Eldin

---

[80] Ibid.

[81] Sam Biddle, "Instagram Hid a Comment. It was Just Three Palestinian Flag Emojis," *The Intercept*, October 28 2023, https://theintercept.com/2023/10/28/instagram-palestinian-flag-emoji/ (accessed November 30, 2023).

[82] See full analysis in section below.

posts frequently about Palestine.[83] He said that he was not able to access the tool that allows him to see potential account violations, and that other users, when trying to tag him in a post, received a warning message that his account had repeatedly posted false information or contravened Community Guidelines.[84]

Other accounts, including the Instagram account of Let's Talk Palestine, which posts educational content about Palestine, reported being temporarily suspended.[85] Meta said, "These accounts were initially locked for security reasons after signs of compromise, and we're working to make contact with the account owners to make sure they have access." The Palestine-based Quds News Network reported that its Facebook page was permanently deleted[86] and that its Instagram account was suspended.[87] Mondoweiss correspondent Leila Warah, who is based in the West Bank, reported in October that Instagram suspended her account. After Mondoweiss publicized the suspension, her account was quickly reinstated, then soon after suspended again and reinstated the following day.[88]

Criticism of Israel as "Hate Speech" and "Dangerous"

Many users reported posts on Instagram being removed when they criticized the Israeli government, including the leadership of Prime Minister Benjamin Netanyahu, no matter how nuanced or careful their posts were. Meta removed these posts under its "Dangerous Organizations or Individuals" and hate speech rules, respectively.

In addition, multiple accounts sharing educational material about Hamas and background information on Palestinian human rights were removed under Meta's DOI policy.[89] Human

---

[83] @ASE, tweet, *X*, December 8, 2023, https://twitter.com/ase/status/1733158080907464817?s=12&t=UcIn9dfSFOEhnqhDWB21iA (accessed December 9, 2023)

[84] ahmedeldin, post, *Instagram*, November 18, 2023, https://www.instagram.com/p/CzywCqtodl6/?igshid=MzRlODBiNWFlZA%3D%3D&img_index=2 (accessed November 30, 2023).

[85] Hibaq Farah, "Pro-Palestinian Instagram account locked by Meta for 'security reasons'", *The Guardian*, October 26, 2023, https://www.theguardian.com/technology/2023/oct/26/pro-palestinian-instagram-account-locked-by-meta-for-security-reasons (accessed November 30, 2023).

[86] @qudsn, tweet with image, *X*, October 14, 2023, https://twitter.com/qudsn/status/1713378047049482681?s=20 (accessed November 30, 2023).

[87] @QudsNen, tweet with image, *X*, November 27, 2023, https://x.com/QudsNen/status/1729259729824764414?s=20 (accessed November 30, 2023).

[88] Prem Thakker, Sam Biddle, "TikTok, Instagram Target Outlet Covering Israel-Palestine Amid Siege on Gaza," *The Intercept*, October 11, 2023, https://theintercept.com/2023/10/11/tiktok-instagram-israel-palestine/ (accessed November 30, 2023).

[89] Evidence ID P56HV, posted to Instagram on October 20, 2023.

Rights Watch reviews found that these posts did not praise or support Hamas but instead were aimed at giving people context and information to understand the escalation in violence.

Human Rights Watch's Call for Censorship Evidence

Dozens of users reported being unable to repost, like, or comment on Human Rights Watch's post calling for evidence of online censorship, which was marked as "spam" and in some cases, flagged under DOI. For example, an account posted about Human Rights Watch's call for censorship documentation and included an email address to send us evidence. Instagram removed the comment, citing a violation of its Community Guidelines.

"Shadow Banned"

While "shadow banning," a type of restriction reported by several hundred users, is challenging to verify, partly due to the lack of platform notice of its occurrence, some users demonstrated compelling evidence to support their claim.[90] This included "before" and "after" screenshots noting the dramatic decrease in number of views after the user started posting content about Palestine, screenshots of engagement metrics, such as likes, comments, and shares, noting a sudden and significant decrease in engagement on content related to Palestine, screenshots that the account or content does not appear in search results, a significant slowdown in new followers, and demonstrating that the content is not visible to others.

Harmful Content that Remained Online

While content that remained online is outside the scope of our research, many users recorded evidence of anti-Palestinian and Islamophobic content that remained online even after they reported it to Instagram and Facebook, in the same post where the users' initial comment was removed. For example, a user reported a comment on their post which said, "Make Gaza a parking lot."[91] After the complaint was reviewed by Instagram, the platform

---

[90] Evidence ID ZW3T8, downranking activity on Instagram on October 22, 2023; Evidence ID 77F87 downranking activity on Instagram on October 27, 2023; Evidence ID Z03H6 downranking activity on Instagram on October 31, 2023; Evidence ID JM0E7 downranking activity on Instagram on October 31, 2023; Evidence ID W49Z6 downranking activity on Instagram on November 3, 2023.

[91] Evidence ID shows that a post consisting in its entirety of the comment "Make GAZA a parking lot." on Instagram was reported on October 22, 2023, and found by Instagram to be consistent with its community guidelines on October 24, 2023.

notified the user that the comment was not removed because it "did not violate Community Guidelines." Another user reported a comment that said, "I wish Israel success in this war in which it is right, I hope it will wipe Palestine off the face of the earth and the map."[92] Instagram found that this post did not violate its Community Guidelines. Another comment, which remained online after being reported, stated, "Imagine an Islamic extremist terrorist accusing us of fascism…lol. Fuck Islam and fuck you. You and your people have done enough to make the world a shittier place for decades."[93]

Underlying Systemic Contributors to Meta's Censorship

A "Dangerous" Policy for Public Debate

Human Rights Watch documented hundreds of cases where Meta applied the DOI policy with the effect of suppressing peaceful speech on issues related to hostilities between Israeli forces and Palestinian armed groups.[94] Because the peaceful content was erroneously restricted, the standard Meta responses were inevitably disproportionate.

Human rights and digital rights organizations have repeatedly highlighted the role the DOI policy plays in silencing Palestinian voices.[95] The UN special rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism expressed concern[96] that the policy is inconsistent with international human rights law,

---

[92] Evidence ID D789L, shows that a post consisting entirely of the comment "I wish Israel success in this war in which it is right, I hope it will wipe Palestine off the face of the earth and the map." Was reported to Instagram on October 26, 2023, and found by Instagram to be consistent with its community guidelines.

[93] Evidence ID 3NXC0, shows that a post consistent in its entirety of "Imagine an islamic extremist terrorist accusing others of fascism…lol. Fuck islam, and fuck you. You and your people have done enough to make the world a shittier place for decades. Stay in your ismalist shithole countries and stay away from the rest of us." was reported to Instagram on November 24, 2023, and was found not to violate Instagram's community guidelines.

[94] Evidence ID AVU7K, showing a story reposting a news videos of hostages being released posted on October 24, 2023 to Instagram are removed by Instagram for violating its guidelines on dangerous organizations and individuals; Evidence ID 4S6ON, showing a comment saying "All that killing for a handful of Hamas leaders? 5,000 innocent dead. Starving and traumatizing a million more. No accuracy huh? Just barbaric, animalistic appetite for murder. War criminals" was removed on October 31, 2023 for violating Instagram's guidelines on violence or dangerous organizations and an appeal was refused, confirming that the post violated the guidelines on dangerous organizations and individuals; Evidence ID 16GA8, showing a reposted image containing a statement in Arabic from the Gaza Ministry of Health indicating that Israeli forces have asked the Ranteesi hospital for children to be evacuated was posted on November 7, 2023, to Instagram, is removed by Instagram for violating its guidelines on dangerous organizations and individuals.

[95] Electronic Frontier Foundation, "Tell Facebook: Stop Silencing Palestine," undated, https://stopsilencingpalestine.com (accessed December 1, 2023).

[96] Mandate of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Communication: OL OTH 46/2018, July 24, 2018, https://www.ohchr.org/sites/default/files/Documents/Issues/Terrorism/OL_OTH_46_2018.pdf (accessed November 30, 2023).

including the rights to free expression, association, participation in political affairs, and non-discrimination.[97] Similarly, the UN special rapporteur on freedom of opinion and expression warned that "Company prohibitions of threatening or promoting terrorism, supporting or praising leaders of dangerous organizations and content that promotes terrorist acts or incites violence are, like counter-terrorism legislation, excessively vague."[98] Even Meta's own Oversight Board has recommended that the company make changes to the policy to avoid censoring protected speech.[99]

The problems with the DOI policy are multilayered, and include how the list is composed, what is covered in the policy, and its enforcement. As noted earlier, because Meta's designation of individuals and entities under the DOI policy relies heavily on US terrorist lists, including its "foreign terrorist organizations" list,[100] it includes political movements that also have armed wings, such as Hamas and the Popular Front for the Liberation of Palestine.[101] It does this even though, as far as is publicly known, US law does not prohibit groups on the list from using free and freely available social media platforms, and does not consider allowing groups on the list to use platforms tantamount to "providing material support" in violation of US law.[102] Meta's adoption of broad and sweeping US designations not only effectively prohibits even peaceful expression of support for many major Palestinian political movements, but prohibits many more Palestinians, including civil

---

[97] Fionnuala Ní Aoláin., "Input of the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism to the Facebook Oversight Board Concerning its 'Community Guidelines' and 'Community Standard on Dangerous Individuals and Organizations'," 2021, https://www.ohchr.org/sites/default/files/Documents/Issues/Terrorism/UNSRCT_Facebook_Oversight_Board_Input2021.docx (accessed November 30, 2023).

[98] UN Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, A/HRC/38/35, April 6, 2018, https://undocs.org/A/HRC/38/35 (accessed November 30, 2023), para. 32.

[99] 2021-009-FB-UA Shared Al Jazeera post, Oversight Board, Overturned, 2021, https://www.oversightboard.com/decision/FB-P93JPXo2 (accessed November 30, 2023); 2021-006-IG-UA Öcalan's isolation, Oversight Board, Overturned, 2021, https://www.oversightboard.com/decision/IG-I9DP23IB, (accessed November 30, 2023); 2022-005-FB-UA Mention of the Taliban in news reporting, Oversight Board, Overturned, 2022, https://www.oversightboard.com/decision/FB-U2HHA647 (accessed November 30, 2023).

[100] U.S. Department of State, "Designated Foreign Terrorist Organizations," https://www.state.gov/foreign-terrorist-organizations/ (accessed December 9, 2023).

[101] Meta, "Facebook Community Standards: Dangerous Organizations and Individuals," https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed November 30, 2023).

[102] Electronic Frontier Foundation, Syrian Archive, and Witness, "Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content," May 2019 https://mnemonic.org/en/content-moderation/impact-extremist-human-rights (accessed December 2, 2023).

servants who work for the local government in Gaza, which Hamas dominates, from using its platforms.

The BSR report found that Palestinians are more likely to violate Meta's DOI policy because of Hamas' presence as a governing entity in Gaza and political candidates' affiliations with designated organizations.[103]

Civil society and the Oversight Board recommended that Meta make public the list of organizations and entities it has designated as dangerous, but Meta has refused to do so, citing employee safety and a concern that doing so would permit banned entities to circumvent the policy. *The Intercept* published a leaked version of the list in October 2021.[104]

The DOI policy not only prohibits "representation," or creating accounts on behalf of designated groups or individuals, but also bans both "praise" and "substantive support," vague and broad terms that include protected expression under international human rights law. For example, Meta defines "praise" as including "speak[ing] positively about a designated entity," giving them "a sense of achievement," legitimizing their cause "by making claims that their hateful, violent, or criminal conduct is legally, morally, or otherwise justified or acceptable," or aligning oneself ideologically with them. Meta defines "substantive support" as directly quoting a designated entity without caption that condemns, neutrally discusses, or is a part of news reporting. The policy recognizes that "users may share content that includes references to designated dangerous organizations and individuals in the context of social and political discourse" and allows for content that reports on, neutrally discusses, or condemns organizations and individuals on the DOI list.

However, if a user's intention is ambiguous or unclear, Meta defaults to removing content. Internal guidance makes this default intent presumption even more problematic by shifting the focus to how content might be perceived rather than on what the user intends. It instructs reviewers to remove content for praise if it "makes people think more positively about" a designated group, making the meaning of "praise" less about the intent of the

---

[103] Business for Social Responsibility (BSR), "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (accessed November 30, 2023).

[104] Sam Biddle, "Revealed: Facebook's Secret Blacklist of 'Dangerous Individuals and Organizations'," *The Intercept*, October 12, 2021, https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/ (accessed November 30, 2023).

speaker but the effects on the audience.[105] The Oversight Board has also criticized Meta's tendency to adjust this policy in secret and make exceptions on an ad hoc basis, for example to discuss conditions of an incarcerated person on the DOI list or to enable people to speak favorably about a positive development from a listed organization acting in a governing capacity.[106] The broad categories of speech covered in the DOI policy combined with the default to removing content if the intent is unclear result in the over-removal of content that should be considered protected speech, even where contextual cues make clear the post is, in fact, reporting.

Violating the DOI policy results in severe penalties on accounts, such as immediate loss of features like live-streaming for 30 days and the ability to have content viewed or recommended to non-followers.[107] Other policy violations would only result in a strike against an account, whereas the penalties for violating the DOI policy are swift and severe.[108] The BSR report noted, "DOI violations also come with particularly steep penalties, which means Palestinians are more likely to face steeper consequences for both correct and incorrect enforcement of policy. In contrast to Israelis and others, Palestinians are prevented from sharing types of political content because the Meta DOI policy has no exemption for the praise of designated entities in their governing capacity."[109]

Human Rights Watch documented the DOI policy being invoked to censor "social and political discourse" around the hostilities, including reporting, neutral discussion, and condemnation of "dangerous" organizations—the type of content the revised DOI policy purports to permit.[110] In one instance, Instagram removed reposting of content from an Arabic language account from the Turkish news broadcast (TRT Arabi) that included a

---

[105] 2022-005-FB-UA Mention of the Taliban in news reporting Oversight Board, Overturned, 2022, https://www.oversightboard.com/decision/FB-U2HHA647 (accessed December 3, 2023).

[106] 2021-006-IG-UA Öcalan's isolation, Oversight Board, Overturned, 2021, https://www.oversightboard.com/decision/IG-I9DP23IB, (accessed November 30, 2023); 2022-005-FB-UA Mention of the Taliban in news reporting Oversight Board, Overturned, 2022, https://www.oversightboard.com/decision/FB-U2HHA647 (accessed December 3, 2023).

[107] Meta, "Protecting Facebook Live From Abuse and Investing in Manipulated Media Research," May 14, 2019, https://about.fb.com/news/2019/05/protecting-live-from-abuse/ (accessed November 30, 2023).

[108] Meta, "Restricting accounts," https://transparency.fb.com/enforcement/taking-action/restricting-accounts/ (accessed November 30, 2023).

[109] Business for Social Responsibility (BSR), "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (accessed November 30, 2023).

[110] Meta, "Facebook Community Standards: Dangerous Organizations and Individuals," https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed November 30, 2023).

statement in Arabic from the Ministry of Health in Gaza that Israeli forces had ordered the Rantisi hospital for children to evacuate before they bombed it. The post was removed for violating Meta's DOI policy, presumably because the Ministry of Health in Gaza is part of a government led by Hamas. Human Rights Watch reviewed more than 100 screenshots documenting removal of Instagram content reposting videos, including from news organizations such as *Middle East Eye*, *Vice News*, and *Al Jazeera* that included reporting on videos of hostages published by Hamas and Islamic Jihad on the basis of DOI policy violations.

The practice by Hamas and Islamic Jihad of publicly releasing videos of hostages constitutes an outrage upon personal dignity, a serious violation of the laws of war.[111] However, Meta adjusted its policy in the weeks following the October 7 attacks to allow hostage imagery when it condemns the act, or which includes information for awareness-raising purposes. The same exceptions apply to any Hamas-produced footage.[112] Prohibiting people from sharing the same videos that news outlets shared without adding language that could reasonably be construed as incitement to violence or hatred hinders the public's ability to engage on issues relating to the crisis. Meta told Human Rights Watch in December that its "teams are considering context around [hostage] imagery, and newsworthy allowances are available where appropriate to balance the public interest against the risk of harm."[113]

Inconsistent and Opaque Prohibitions on Newsworthy Content

Meta platforms host images, videos, and posts from news outlets, independent journalists, and other sources from conflict zones. At times, this media may include violent and graphic content, hate speech, or nudity. Although Meta policy prohibits violent and graphic content,[114] hate speech,[115] violence and incitement,[116] and nudity and sexual

---

[111] "Gaza: Hostage Videos an 'Outrage on Personal Dignity,'" Human Rights Watch news release, November 10, 2023, https://www.hrw.org/news/2023/11/10/gaza-hostage-videos-outrage-personal-dignity.

[112] Letter from Miranda Sissons, Director of Human Rights Policy, Meta, to Human Rights Watch, December 6, 2023.

[113] Ibid.

[114] Meta, "Facebook Community Standards: Violent and Graphic Content," https://transparency.fb.com/policies/community-standards/violent-graphic-content/ (accessed November 30, 2023).

[115] Meta, "Facebook Community Standards: Hate Speech," https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed November 30, 2023).

[116] *Meta*, "Facebook Community Standards: Violence and Incitement,"https://transparency.fb.com/policies/community-standards/violence-incitement/ (accessed November 30, 2023).

activity,[117] the company makes an exception[118] if it deems the content to be newsworthy and in the service of public interest.

Meta uses a post depicting violence in Ukraine as an illustrative example of the importance of the newsworthiness allowance,[119] demonstrating the company's willingness to adjust its policies to account for the realities of conflict in another high-profile conflict. When properly enforced, Meta's newsworthiness allowance has the capacity to bolster discourse, raise public awareness, and facilitate research, including that done by human rights investigators.[120]

However, Human Rights Watch's investigation found that Meta has inconsistently enforced its newsworthiness allowance policy and has misapplied its prohibitions on incitement and nudity to newsworthy content that does not appear to violate those policies. More specifically, the research shows Meta platforms have repeatedly removed media of graphic content from Palestine, effectively censoring such images.[121] This media includes photos of injured and murdered Palestinians, a video of Israelis urinating on Palestinians, and a Palestinian child shouting "Where are the Arabs?" after his sister was killed.[122] In these cases, content was removed for violating Meta's policy on violence and incitement. In these cases, the news value of the shared material was such that it is hard to justify a decision to block this content on the basis of a policy on violence and incitement.

---

[117] Meta, "Facebook Community Standards: Adult Nudity and Sexual Activity," https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/ (accessed November 30, 2023).

[118] Meta, "Our approach to newsworthy content," August 29, 2023, https://transparency.fb.com/features/approach-to-newsworthy-content/ (accessed November 30, 2023).

[119] Ibid. The example reads: Post by Ukrainian Defense Ministry depicting charred bodies "This video originally shared by the Ukrainian Defense Ministry very briefly depicts an unidentified charred body. Though we typically remove this type of content under our Violent and Graphic Content policy, we determined that this video qualified for a newsworthy allowance, as it documented an ongoing armed conflict. We placed a warning screen over this content and limited its availability to adults ages 18 and older because of the graphic nature of the content."

[120] Human Rights Watch, *"Video Unavailable" Social Media Platforms Remove Evidence of War Crimes*, (New York: Human Rights Watch, 2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.

[121] "Russia, Ukraine, and Social Media and Messaging Apps: Questions and Answers on Platform Accountability and Human Rights Responsibilities," Human Rights Watch questions and answers piece, March 16, 2022, https://www.hrw.org/news/2022/03/16/russia-ukraine-and-social-media-and-messaging-apps.

[122] Evidence ID BA5X0 shows a video of a child yelling "where are the Arabs" after his sister was killed that was posted to Instagram on October 21, 2023, and removed by Instagram for violating its guidelines on violence and incitement. Evidence ID oKE5Y shows a video of Israelis urinating on deceased Palestinians and kicking their bodies that was posted to Instagram on October 21, 2023, and removed for violating its guidelines on violence and incitement.

Five Instagram users and one Facebook user reported that images of injured and dead bodies in Gaza's hospitals were removed for violating the platform's Community Guidelines regarding violence and incitement. Meta's violence and incitement guidelines prohibit "language that incites or facilitates serious violence" with the stated intention of preventing offline harm. The six images removed made no call for violence.[123]

Additionally, multiple users reported that Instagram removed content depicting the plight of Palestinians, ostensibly for violating its nudity or sexual activity policy. This content includes images of killed Palestinians, a video that appears to show IDF soldiers humiliating and torturing Palestinians, and an image of bombings on Gaza. Three users reported that an image of a fully clothed man holding a girl, both deceased, was removed for violating the platform's policy on nudity or sexual activity. Instagram removed this image even though it did not include any nudity or sexual activity and likely met the newsworthiness allowance in Meta's own guidelines.

Meta's failure to apply the newsworthiness allowance to this content not only functions to censor images of abuse of Palestinians, but also suggests that Meta does not consider such images to serve the public interest.

Some users who reported cases to Human Rights Watch explained that their posts sought to speak out against violence, not incite it. By stripping the content of context and bluntly applying its policies, Meta is effectively censoring newsworthy content and achieving the opposite outcome of the stated intention of its policies.

Where Meta highlighted the importance of the newsworthiness allowance as it applied to Ukraine-related content that it might otherwise prohibit, it appears to have failed to extend the same policy to content documenting the impact of the current hostilities on Palestinians. Far from recognizing the heightened need for latitude in applying its content

---

[123] Evidence ID 5O567, comment and video posted to Facebook on October 15, 2023, showing a child saying goodbye to deceased mother with text added by a user saying: "no mother no father for many little ones a Palestinan child bids farewell to his…" is removed by Facebook on October 17, 2023; Evidence ID 99OU8, story posted to Instagram on October 17, 2023, showing doctors holding a press conference with dead children is removed by Instagram for violating its policy on violence and incitement on October 17, 2023; Evidence ID M8LZ3, a comment and image discussing the bombing of Baptist Hospital in Gaza City posted to Instagram on October 17, 2023, is removed by Instagram for violating its policies on violence and incitement; Evidence ID 89X5O showing an image of a dead child in a hospital with text added that reads: "hello world thank you for watching", posted October 19, 2023, to Instagram and removed by Instagram for violating its policy on violence and incitement.

prohibitions in discussions of ongoing hostilities, the examples shared with Human Rights Watch suggest Meta is applying community guidelines aggressively to content that should not be prohibited in the first place. The suppression by Meta platforms of content documenting Palestinian injury and death can result in offline harm, as gaps in information impact public understanding and resulting political responses.

Lack of Transparency Around Government Requests

Meta removes content based on its Community Standards[124] and to comply with local laws.[125] The company regularly reports on both types of content restrictions.[126]

However, Meta takes down a significant amount of content in response to requests by governments for "voluntary" takedown based on alleged violations of the company's Community Standards. Such requests come from internet referral units (IRUs),[127] which vary by country but are generally non-judicial bodies, like law enforcement authorities or administrative bodies.

IRU requests typically risk circumventing legal procedures, lack transparency and accountability, and fail to provide users with access to effective remedy. They deny people the due process rights they would have if the government sought to restrict the content through legal processes. Unlike content takedowns based on local law, which should be based on legal orders and result in geolocated restrictions on content, takedowns based on Meta's Community Standards result in removal of that content globally. Furthermore, the user is not notified that the removal of their content is due to a government request, nor is the role of the government reflected in Meta's biannual transparency reports.

The Israeli government has been aggressive in seeking to remove content from social media. The Israeli Cyber Unit, based within the State Attorney's Office, flags and

---

[124] Meta, "Community Standards Enforcement Report," https://transparency.fb.com/reports/community-standards-enforcement/ (accessed November 30, 2023).

[125] Meta, "How we assess reports of content violating local law," https://transparency.fb.com/reports/content-restrictions/content-violating-local-law (accessed November 30, 2023).

[126] Meta, "Content Restrictions Based on Local Law: Israel," https://transparency.fb.com/reports/content-restrictions/country/IL (accessed November 30, 2023).

[127] Global Network Initiative, "Understanding the Human Rights Risks Associated with Internet Referral Units." February 25, 2019, https://globalnetworkinitiative.org/human-rights-risks-irus-eu/ (accessed November 30, 2023).

submits requests to social media companies to "voluntarily" remove content.[128] Instead of going through the legal process of filing a court order based on Israeli criminal law to take down online content, the Cyber Unit makes appeals directly to platforms based on their own terms of service. Since Israel's State Attorney's Office began reporting on the Cyber Unit's activities, platforms' overall compliance rate with its requests has never dropped below 77 percent and in 2018 was reported to be as high as 92 percent.[129]

Requests from the Cyber Unit to Meta platforms are far higher than what Meta reports as legal removal orders from the Israeli government. In 2021, the Cyber Unit issued 5,990 content removal or restriction requests, with an 82-percent compliance rate across all platforms.[130] The majority of requests (around 90 percent) were directed to Facebook and Instagram and were issued during the escalation of hostilities in May 2021. That same year Meta reported that it had restricted 291 pieces of content or accounts on Facebook and Instagram based on local law in response to government in Israel.[131]

According to media reports on November 14, 2023, the prosecutor's office has sent 9,500 content takedown requests since October 7, 2023, to major social media platforms related

---

[128] Adalah, "Israel State Attorney claims censorship of social media content, following Cyber Unit requests, isn't an 'exercise of gov't authority'," November 28, 2019, https://www.adalah.org/en/content/view/9859 (accessed November 30, 2023).

[129] Ministry of Justice, Office of the State Attorney, Annual Report, 2021, https://www.gov.il/BlobFolder/news/report2021/he/2021-year-report.pdf (accessed December 12, 2023), page 78, "82% of requests led to the removal of reported content"; Ministry of Justice, Office of the State Attorney, Annual Report, 2020,
https://www.gov.il/BlobFolder/reports/office_of_the_state_2020/he/office_of_the_state_2020.pdf (accessed December 12, 2023), compliance rate not reported; Ministry of Justice, Office of the State Attorney, Annual Report, 2019,
https://www.gov.il/BlobFolder/generalpage/files-general/he/DATA%202019.pdf (accessed December 12, 2023), Figure 51 (90% removed, 1% partially removed, 9% not removed); Ministry of Justice, Office of the State Attorney, Annual Report, 2018,
https://www.gov.il/BlobFolder/generalpage/files-general/he/files_report-2018.pdf (accessed December 12, 2023), Figure 55 (86% fully removed; 6% partially removed; 8% content not removed); Ministry of Justice, Office of the State Attorney, Annual Report, 2017, https://www.gov.il/BlobFolder/reports/annual-report-2017/he/files_data-report-2017.pdf (accessed December 12, 2023), Figure 47, 85% fully removed, 3% partially removed, 12% not removed; Ministry of Justice, Office of the State Attorney, Cyber Unit, Annual Report, 2016, https://web.archive.org/web/20220119190236/https://www.justice.gov.il/Units/StateAttorney/Documents/annualcyber.pdf (accessed December 12, 2023), p 5: in its inaugural 2016 annual report, the Cyber Unit states that since it began operations it had issued 2,138 content removal requests social media companies that had been resolved. Of these 1,716 (or 80.3%) had resulted in partial (162) or complete (1,554) content removal while 422 (19.7%) resulted in no content being removed. An additional 52 content removal requests are reported as being still under consideration. This 2016 report describes each of the 2,138 content removal requests as potentially including dozens or even hundreds of pieces of actual allegedly infringing content (posts, videos, etc.).

[130] Ministry of Justice, Office of the State Attorney, Annual Report, 2021,
https://www.gov.il/BlobFolder/news/report2021/he/2021-year-report.pdf (accessed December 12, 2023).

[131] Meta, "Transparency Center: Content Restrictions Based on Local Law: Israel",
https://transparency.fb.com/reports/content-restrictions/country/IL/ (accessed December 12, 2023), "Amount of content we restricted": Meta reports 267 items of content restricted in Jan-June 2021 and 24 items of content restricted in July-December 2021.

to the recent hostilities that they allege violate the companies' policies.[132] Nearly 60 percent of those requests went to Meta. Media reports cite a 94-percent compliance rate for such requests across platforms.[133] Human Rights Watch inquired with the Cyber Unit what company policy the post or account allegedly violated but did not receive a response at time of writing.

IRUs from other countries may also be requesting that Meta and other platforms remove content about the hostilities in Israel and Gaza. The European Commissioner for Internal Market, Thierry Breton, recently sent letters to the heads of major social media platforms, including Meta CEO Mark Zuckerberg, about an increase in "illegal content and disinformation being disseminated in the EU" following the "terrorist attacks carried out by Hamas against Israel." The letter requested that Meta be "very vigilant to ensure strict compliance with the [Digital Services Act (DSA)] rules on terms of service, on the requirement of timely, diligent and objective action following notices of illegal content in the EU, and on the need for proportionate and effective mitigation measures."[134]

While 30 digital rights organizations questioned Breton's interpretation of the DSA contained in the letter,[135] the DSA does provide for the establishment of "trusted flaggers," to notify platforms about illegal content, notifications that should be processed and decided upon with priority and without delay.[136] The DSA explicitly says that law enforcement agencies can be designated as "trusted flaggers."[137] While their notices

---

[132] Thomas Brewster, "Israel Has Asked Meta And TikTok To Remove 8,000 Posts Related To Hamas War," *Forbes,* November 14, 2023, https://www.forbes.com/sites/thomasbrewster/2023/11/13/meta-and-tiktok-told-to-remove-8000-pro-hamas-posts-by-israel/?sh=16af3f16f6ce (accessed November 30, 2023).

[133] Ibid.

[134] Letter from Thierry Breton to Mark Zuckerberg, October 11, 2023, posted on Twitter, https://twitter.com/thierrybreton/status/1712126600873931150?s=46&t=Y-CDvNYEVAdPCdphstKDgQ (accessed November 30, 2023).

[135] Access Now, "Precise interpretation of the DSA matters especially when people's lives are at risk in Gaza and Israel," October 18, 2023, https://www.accessnow.org/press-release/precise-interpretation-of-dsa-matters-in-gaza-and-israel/ (accessed November 30, 2023).

[136] European Union, "Regulation (EU) 2022/2065 of the European Parliament and of the Council 19 October on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance)," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065 (accessed November 30, 2023).

[137] Ibid.

merely allege illegal content, platforms are likely to treat them as an order to remove the content given the significant legal risk they would face in failing to act.[138]

Echoing civil society, the Oversight Board has expressed concern that users whose content is removed based on Community Standards should be informed where there was government involvement in content removal. In an unrelated case, the Oversight Board recommended that Meta ensure users are notified when their content is removed due to a government request under Community Standards violations and that it ensure a transparent process for receiving and responding to all government requests for content removal.[139] Further, it recommended that Meta include information on the number of requests it receives for content removals from governments that are based on Community Standards violations (as opposed to violations of national law), and the outcome of those requests. In August 2021, Meta said it was fully implementing these recommendations.[140]

In the "Shared Al Jazeera post" case,[141] the Board again recommended that Meta improve transparency around government requests that led to global removals based on violations of the company's Community Standards; in October 2021, Meta said it was implementing this recommendation in part.[142] The BSR report also recommended that Meta disclose the number of formal reports received from government entities about content that is not illegal, but which potentially violates Meta content policies.[143]

---

[138] Center for Democracy and Technology, "A Series on the EU Digital Services Act: Tackling Illegal Content Online," August 2, 2022, https://cdt.org/insights/a-series-on-the-eu-digital-services-act-tackling-illegal-content-online/ (accessed November 30, 2023)

[139] 2021-006-IG-UA Öcalan's isolation, Oversight Board, Overturned, 2021, https://www.oversightboard.com/decision/IG-I9DP23IB, (accessed November 30, 2023).

[140] Meta, "Case regarding the support of Abdullah Öcalan, founder of the PKK," June 12, 2023, https://transparency.fb.com/oversight/oversight-board-cases/support-of-abdullah-ocalan-founder-of-the-pkk/ (accessed November 30, 2023)

[141] 2021-009-FB-UA Shared Al Jazeera post, (2021). Oversight Board, Overturned, 2021, https://www.oversightboard.com/decision/FB-P93JPX02https://www.oversightboard.com/decision/FB-P93JPX02 (accessed November 30, 2023)

[142] Meta, "Oversight Board Selects a case related to an Al Jazeera post on tensions between Israel and Palestine," June 12, 2023, https://transparency.fb.com/en-gb/oversight/oversight-board-cases/al-jazeera-post-tensions-israel-palestine (accessed November 30, 2023).

[143] Business for Social Responsibility (BSR), "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (accessed November 30, 2023).

Meta's September 2023 status update on its implementation of BSR's recommendation says its efforts in this area are in-progress, which it describes as "a complex, long-term project."[144] Meta said it would "provide an update on the timeline for public reporting of these metrics in a future Oversight Board Quarterly Update and in [its] next annual Human Rights Report." More than two years after committing to publishing data around government requests for taking down content that is not necessarily illegal, Meta has failed to increase transparency in this area.

Reliance on Automation

Meta's reliance on automation for content moderation is a significant factor in the erroneous enforcement of its policies, which has resulted in the removal of non-violative content in support of Palestine on Instagram and Facebook.

According to Meta, over 90 per cent of the content deemed to violate their policies is proactively detected by their automated tools before anyone reports it.[145] Automated content moderation is notoriously poor at interpreting contextual factors that can be key to determining whether a post constitutes support for or glorification of terrorism. This can lead to overbroad limits on speech and improper labeling of it as violent, criminal, or abusive.[146] [147]

Meta relies on automation to detect and remove content deemed violative by the relevant platform and reposting of that content, regardless of complaints. It also uses algorithms to determine which automated removals should be prioritized for human oversight, as well as

---

[144] Meta, "Meta Update: Israel and Palestine Human Rights Due Diligence," September 2023, https://humanrights.fb.com/wp-content/uploads/2023/09/September-2023-Israel-and-Palestine-HRDD-Meta-Update.pdf (accessed November 30, 2023).

[145] Meta, "How technology detects violations," October 18, 2023, https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations (accessed November 30, 2023).

[146] Center for Democracy and Technology, "Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis," May 2021, https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf (accessed November 30, 2023).

[147] Automated moderation of content that platforms consider to be "terrorist and violent extremist" has in other contexts led to the removal of evidence of war crimes and human rights atrocities from social media platforms in some cases before investigators are aware that the potential evidence exists, See: Human Rights Watch, "Video Unavailable" Social Media Platforms Remove Evidence of War Crimes, (New York: Human Rights Watch, 2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes. Electronic Frontier Foundation, Syrian Archive, and Witness, "Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content," May 2019, https://mnemonic.org/en/content-moderation/impact-extremist-human-rights (accessed November 30, 2023).

in processing existing complaints and appeals.[148] Meta reported on October 13, 2023, that it was taking temporary steps to lower the threshold at which it takes action against potentially violating and borderline content across Instagram and Facebook,[149] to avoid recommending this type of content to users in their feeds.[150] However, these measures increase the margin of error and result in false positives flagging non-violative content.

Meta does not publish data on automation error rates or on the degree to which automation plays a role in processing complaints and appeals. Meta's lack of transparency hinders the ability of independent human rights and other researchers to hold its platforms accountable, allowing wrongful content takedowns as well as ineffective moderation processes for violative content to remain unchecked. Processes intended to remove extremist content, in particular the use of automated tools, have sometimes perversely led to removing speech opposed to terrorism, including satire, journalistic material, and other content that would, under rights-respecting legal frameworks, be considered protected speech.[151]

In reviewing hundreds of cases of content removal and the inability of certain users to post comments on Instagram and Facebook, Human Rights Watch found Meta's automated moderation tools failed to accurately distinguish between peaceful and violent comments. Users reported that their ability to express opinions, including dissenting or unpopular views about the escalation of violence since October 7, is being restricted repeatedly and increasingly over time. As a result of comment removal or restriction, users reported altering their behavior on Instagram and Facebook to avoid their comments being removed. After multiple experiences with seemingly automated comment removal, users reported being less likely to engage with content, express their opinions, or participate in discussions about Israel and Palestine.

[148] Meta, "How We Review Content," August 11, 2020, https://about.fb.com/news/2020/08/how-we-review-content/ (accessed December 13, 2023).

[149] Meta, "Meta's Ongoing Efforts Regarding the Israel-Hamas War," October 13, 2023 (Updated October 18, 2023), https://about.fb.com/news/2023/10/metas-efforts-regarding-israel-hamas-war/ (accessed November 30, 2023).

[150] Meta, "Recommendation Guidelines," August 31, 2020, https://about.fb.com/news/2020/08/recommendation-guidelines/ (accessed November 30, 2023).

[151] "Joint Letter to New Executive Director, Global Internet Forum to Counter Terrorism," Human Rights Watch, July 30, 2020, https://www.hrw.org/news/2020/07/30/joint-letter-new-executive-director-global-internet-forum-counter-terrorism.

Human Rights Implications of Palestine Content Censorship

Content Restrictions and "Shadow Banning"

Article 19 of the International Covenant on Civil and Political Rights (ICCPR)[152] guarantees the right to freedom of expression, including the right to seek, receive, and impart information and ideas of all kinds.[153] This right applies to online expression, as the UN Human Rights Committee has clarified.[154]

The right to freedom of expression is not absolute. Limitations on this right are possible if they are necessary for and proportionate to the protection of national security, public order, public health, morals, or the rights and freedoms of others. Limitations for these purposes must be established in law, not impair the essence of these rights, and be consistent with the right to an effective remedy.[155] The same standard applies to limitations of the rights to freedom of assembly and association.[156]

Unduly restricting or suppressing peaceful content that supports Palestine and Palestinians impermissibly infringes on people's rights to freedom of expression. Given that social media has become the digital public sphere and the site of social movements, undue restrictions on content and the ability to engage with other users on social media also undermine the rights to freedom of assembly and association, as well as participation in public affairs. The enforcement of content removal policies and adjustments to recommender algorithms, which determine what content people see in their feeds, to significantly limit circulation of content may be perceived as biased or selectively targeting specific viewpoints and could undermine the right to non-discrimination and the universality of rights as well as the right to due process.

---

[152] International Covenant on Civil and Political Rights (ICCPR), December 16, 1966, 999 U.N.T.S. 171 (entered into force March 23, 1976).

[153] ICCPR, art. 19(2).

[154] UN Human Rights Committee, General Comment No. 34, Freedoms of opinion and expression, CCPR/C/GC/34, 2011, https://undocs.org/en/CCPR/C/GC/34 (accessed December 9, 2023), paras. 12 and 43.

[155] See Human Rights Committee, General Comment No. 34: Article 19: Freedoms of Opinion and Expression, U.N. Doc. CCPR/C/GC/34, September 12, 2011, https://undocs.org/CCPR/C/GC/34, paras. 21-36; Human Rights Committee, General Comment No. 37 on the Right of Peaceful Assembly, U.N. Doc. CCPR/C/GC/37, September 17, 2020, https://undocs.org/CCPR/C/GC/37, paras. 36-49.

[156] ICCPR, arts. 21, 22.

Removing or suppressing online content can hinder the ability of individuals and organizations to advocate for human rights of Palestinians and raise awareness about the situation in Israel and Palestine. Content removal that is carried out using automated systems, such as on Instagram and Facebook, raises concerns about algorithmic bias. As this report documents, these systems may result in the erroneous suppression of content, leading to discriminatory consequences without opportunity for redress.

Engaging with content, such as posting or reading comments, is a crucial aspect of social media interaction, especially when open discussion is prohibited and contested in offline spaces. Being shadow banned—where a user's content is seemingly not visible as usual to their friends and followers, without explanation—can be distressing for users. Meta does not formally acknowledge the practice of shadow banning, effectively denying users transparency, as well as adequate access to complaints mechanisms and meaningful remedy. Social media can be a vital communications tool in crises and conflicts. However, users experiencing or even aware of the risk of account restrictions like shadow banning may refrain from engaging on social platforms in order to avoid losing access to their accounts and vital information, resulting in self-censoring behaviors.

Inability to Appeal to Platform

The UN Guiding Principles on the Business and Human Rights (UNGPs) require businesses to provide access to a remedy where they identify that they have caused or contributed to adverse impacts.[157] This report documents over 300 cases in which users reported and provided evidence of being unable to appeal content removals or account restrictions due to the appeal mechanism malfunctioning. This left them with no effective access to a remedy.

Meta's temporary measures to lower the threshold at which it takes action against potentially violating and borderline content across Instagram and Facebook to avoid recommending this type of content to users in their feeds is likely to increase the margin of error for removal or suppression of content and leave the user without the ability to remedy because the user is not informed of any action taken on their account or content.

---

[157] Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, in Human Rights Council, Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, U.N. Doc. A/HRC/17/31, March 21, 2011, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (accessed November 30, 2023), annex.

Meta told Human Rights Watch that it is aware that the temporary measures it takes during conflicts could have unintended consequences "like inadvertently limiting harmless or even helpful content" and also admitted that"[d]uring busy periods, such as during conflict situations, we may not always be able to review everything based on our review capacity." [158] Meta also disclosed that "appeals for content demotions are currently not available outside of the EU."

The lack of effective remedy for incidents of censorship can have significant implications for individuals and groups. Their right to freedom of expression, as outlined in international human rights instruments, may be violated.

---

[158] Letter from Miranda Sissons, Director of Human Rights Policy, Meta, to Human Rights Watch, December 6, 2023.

III. Social Media Companies' Responsibilities

Under the United Nations Guiding Principles on Business and Human Rights (UNGPs), companies have a responsibility to respect human rights by avoiding infringing on human rights, identifying and addressing the human rights impacts of their operations, and providing meaningful access to a remedy.[159] For social media companies, this responsibility includes aligning their content moderation policies and practices with international human rights standards, ensuring that decisions to take content down are not overly broad or biased, being transparent and accountable in their actions, and enforcing their policies in a consistent manner.

UNGPs require companies to carry out human rights due diligence to identify, prevent, mitigate, and account for how they address their adverse human rights impacts. Companies should communicate externally how they are addressing their human rights impacts, providing sufficient information so that stakeholders can evaluate the adequacy of their response. Meta's Corporate Human Rights Policy outlines its commitment to respecting human rights as set out in the UNGPs.[160] As a member of the Global Network Initiative (GNI),[161] Meta has also committed to upholding the GNI Principles on Freedom of Expression and Privacy.[162]

The Santa Clara Principles on Transparency and Accountability in Content Moderation provide important guidance for how companies should carry out their responsibilities in upholding freedom of expression.[163] Based on those principles, companies should clearly

---

[159] Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, in Human Rights Council, Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, U.N. Doc. A/HRC/17/31, March 21, 2011, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (accessed November 30, 2023), annex.

[160] Meta, "Corporate Human Rights Policy," https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf (accessed November 30, 2023).

[161] The Global Network Initiative, https://globalnetworkinitiative.org/ (accessed November 30, 2023).

[162] The Global Network Initiative, The GNI Principles, https://globalnetworkinitiative.org/gni-principles/ (accessed November 30, 2023).

[163] The Santa Clara Principles on Transparency and Accountability in Content Moderation, 2021, https://santaclaraprinciples.org/ (accessed November 30, 2023).

explain to users why their content or their account has been taken down, including the specific clause of the Community Standards that the content was found to violate.

Companies should also explain how the content was detected, evaluated, and removed—for example, by users, automation, or human content moderators—and provide a meaningful opportunity for timely appeal of any content removal or account suspension. Meta has endorsed the Santa Clara Principles[164] but has not fully applied them.

---

[164] Electronic Frontier Foundation, "Who Has Your Back? Censorship Edition 2019," June 12, 2019, https://www.eff.org/fa/wp/who-has-your-back-2019#santa-clara-principles (accessed November 30, 2023).

IV. Recommendations

To Meta (Instagram and Facebook)

Dangerous Organizations and Individuals (DOI) Policy

- Overhaul the DOI policy so that it is consistent with international human rights standards, in particular, to ensure that Meta platforms permit protected expression, including about human rights abuses, political movements, and organizations that Meta or governments designate as terrorist.
- Instead of relying primarily on a definition of terrorist entities or dangerous organizations, refocus the policy on prohibiting incitement to terrorism, drawing on the model definition advanced by the mandate of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism.[165]
- To the extent that a revised policy includes a definition of terrorist entities or dangerous organizations, do not rely exclusively on the lists of particular states in determining whether to bar an organization.
- Publish Meta's list of Dangerous Organizations and Individuals.
- Clarify which of the organizations banned by Israeli authorities are included under Meta's Dangerous Organizations and Individuals policy.
- Ensure accounts that trigger violations under the DOI policy are subject to proportionate penalties, given the propensity of this policy to erroneously flag protected expression, including about human rights abuses.

Government Requests

- Improve transparency around voluntary government requests to remove content based on Community Standards and Community Guidelines from Israel's Cyber Unit and other internet referral units.

---

[165] UN Human Rights Council, Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Martin Scheinin - Ten areas of best practices in countering terrorism, A/HRC/16/51, December 22, 2010, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/16/51 (accessed November 30, 2023), para. 32.

- o Notify users if a government was involved in their content being taken down based on a policy violation, and provide a transparent appeal process for the decision.
- o Meta should include in its periodic transparency reports:
  - Number of requests per country (broken down by government agency).
  - Compliance rate per country.
  - The relevant company policy the post or account allegedly violated.
  - Compliance rate per policy.

Newsworthiness Allowance

- Conduct an audit to determine error rates concerning the removal of content that is of public interest and should be retained on Meta's platforms under its newsworthiness allowance. This audit should also assess whether Meta is applying the newsworthiness allowance equitably and in a non-discriminatory manner.
- Improve systems to identify and allow listing pieces of content that are newsworthy but are repeatedly removed erroneously.

Automation

- Improve transparency about where and how automation and machine learning algorithms are used to moderate or translate Palestine-related content, including sharing information on the classifiers programmed and used, and their error rates.
- Conduct due diligence to assess the human rights impact of temporary changes in Meta's recommendation algorithms in response to October 7, and share those findings publicly. This assessment and reporting should become standard practice whenever Meta applies temporary measures in crisis situations.
- Integrate the human-in-the-loop principle, wherein humans have a role in the ultimate decision-making process, for meaningful oversight of decisions made by Artificial Intelligence (AI) tools. This is also consistent with the UN Guiding Principles on Business and Human Rights (UNGPs), which require companies to set up internal accountability mechanisms for the implementation of policies and facilitate the right to remedy.

Transparency and Access to Remedy

- Provide users with adequate information when notifying them that their account or content has been restricted, including:
  - The specific content or behavior that violated Meta's Community Guidelines, including the specific clause of the Community Guidelines that their content was found to violate and how the content was detected and removed (for example, flagged by other users of automated detection).
  - The restriction placed on their account or content, including when their account or content has been removed or downgraded in recommender algorithms.
  - How the user can appeal this decision.
- Ensure that all appeal mechanisms are accessible, functional, and available to all users, regardless of jurisdiction.
- Commission and publish an external audit into shadow banning with an aim towards improving public understanding of what changes Meta has made to its recommender systems, content ranking, and the penalty system, and its impact on freedom of expression.

Human Rights Due Diligence

- Solicit feedback from civil society and other relevant stakeholders on Meta's implementation of commitments made in response to the BSR report and the Oversight Board to inform its own assessment of progress made.
- Work with civil society and other relevant stakeholders to timeline implementation of outstanding commitments based on urgency.

Preservation

- Preserve and archive material of human rights violations and abuses that may have evidentiary value, and provide access to data for independent researchers and investigators, including those in the fields of human rights, while protecting user privacy.

Acknowledgments

Human Rights Watch also benefited greatly from expert input from and collaboration with 7amleh, Access Now, and Amnesty International.

Human Rights Watch is grateful for all those who shared their experiences with us.

Annex I: Human Rights Watch Letter to Meta

Human Rights Watch

350 Fifth Avenue, 34th Floor
New York, NY 10118-3299
Tel: +1-212-290-4700
Fax: +1-212-736-1300; 917-591-3452

Tirana Hassan, Executive Director

Deputy Executive Director

Wisla Heneghan, DED/ Chief Operating Officer

Senior Leadership

Sari Bashi, Program Director
Lauren Camilli, General Counsel
Mei Fong, Chief Communications Officer
Abir Ghattas, Chief Information Officer (Acting)
Marusya Lazo, Chief Financial Officer
Joe Lisi, Chief People Officer (Acting)
Valentina Rosa, Chief Development Officer
James Ross, Legal and Policy Director
Bruno Stagno Ugarte, Chief Advocacy Officer
Minjon Tholen, Global Head of Diversity, Equity, and Inclusion

Division and Program Directors

Lama Fakih, Middle East & North Africa
Juanita Goebertus, Americas
Tanya Greene, United States
Elaine Pearson, Asia
Mausi Segun, Africa
Hugh Williamson, Europe and Central Asia

Shantha Rau Barriga, Disability Rights
Elizabeth Evenson, International Justice
Bill Frelick, Refugee and Migrant Rights
Arvind Ganesan, Economic Justice and Rights
Steve Goose, Arms
Maria McFarland Sánchez-Moreno, Technology and Human Rights (Acting)
Zama Neff, Children's Rights
Richard Pearshouse, Environment and Human Rights
Graeme Reid, Lesbian, Gay, Bisexual, and Transgender Rights
Macarena Sáez, Women's Rights
Ida Sawyer, Crisis and Conflict

Advocacy Directors

Yasmine Ahmed, United Kingdom
Maria Laura Canineu, Brazil
Louis Charbonneau, United Nations, New York
Philippe Dam, European Union
Farida Deif, Canada
Kanae Doi, Japan
Meenakshi Ganguly, South Asia
Daniela Gavshon, Australia
Bénédicte Jeannerod, France
Iskra Kirova, Europe and Central Asia
Wenzel Michalski, Germany
Måns Molander, Nordic
Allan Ngari, Africa
Hilary Power, United Nations, Geneva
Sarah Yager, Washington DC

Board of Directors

Amy Rao, Co-Chair
Neil Rimer, Co-Chair
Oki Matsumoto, Vice Chair
Amy Towers, Vice Chair; Treasurer
Catherine Zennström, Vice Chair
Bruce Rabb, Secretary
Akwasi Aidoo
Lishan Aklog
George Coelho
Roberto Dañino
Kimberly Marteau Emerson
Loubna Freih
Leslie Gilbert-Lurie
Paul Gray
Caitlin Heising
David Lakhdhir
Louisa Lee-Reizes
Alicia Miñana
Gloria Principe
Ambassador Robin Sanders
Bruce Simpson
Joseph Skrzynski AO
Donna Slaight, C.M.
Siri Stolt-Nielsen
Marie Warburg
Isabelle de Wismes
Masa Yanagisawa
Andrew Zolli

Nick Clegg
President, Global Affairs
Facebook Inc. (Meta)

Monika Bickert
Vice President, Content Policy
Facebook Inc. (Meta)

November 15, 2023

Dear Mr. Clegg and Ms. Bickert,

We are writing to you on behalf of Human Rights Watch to share the findings of research we have conducted on the censorship and suppression of content related to Israel and Palestine on social media since October 7, and to solicit Meta's perspective by responding to several specific questions that have stemmed from the research.

Human Rights Watch has been documenting online censorship related to Israel and Palestine since October 7 to November 15, a period during which we have observed takedowns or other suppression of content by Palestinians and their supporters, including about human rights abuses, on Instagram and Facebook in significant numbers and even when similar content by other users is not suppressed. In October, Human Rights Watch published a call for evidence of online censorship and suppression of content related to Israel and Palestine on social media since October 7, which we posted from the HRW main accounts on multiple social media platforms, including Instagram, Facebook, X, and TikTok.

We have gathered hundreds of cases which detail various forms of censorship impacting posts and accounts documenting, condemning, and raising awareness about the recent and ongoing hostilities in Israel and Palestine, which involve serious human rights abuses.

Examples include deleting posts, stories, and comments, and suspending accounts, at times without warning the user. We have also documented users' inability to engage with content, such as commenting, sharing, and reposting on stories, on Instagram and Facebook, and restrictions on users' ability to use certain features, such as Instagram/Facebook Live, for prolonged periods. The most widely cited reasons for these restrictions on content and accounts were Community Guidelines or Standards violations, specifically those relating to "dangerous organizations and individuals," "adult nudity and sexual activity," "hate speech," "violent and graphic content," and "spam." In dozens of cases we have documented, users have been unable to appeal content or account removal due to the appeal mechanism malfunctioning, meaning that they have no effective access to a remedy.

While Meta attributed some users' significantly reduced reach to reshared Reels and Feed posts not showing up properly in their Stories on Instagram and inability to go Live on Facebook to a "bug," and reported having fixed the issue on October 18, we have continued to receive cases of such restrictions after that date.

Another common complaint was "shadowbanning," specifically reports of reduction in the distribution or reach of content relating to developments in Israel and Palestine and disabling searches for accounts that post about Palestine. We note that Meta has reported taking steps to lower the threshold at which it takes action against potentially violating and borderline content across Instagram and Facebook to avoid recommending this type of content and to reduce the visibility of potentially offensive comments. We are concerned that these measures increase the margin of error and result in false positives flagging non-violative content. Furthermore, because the user is not informed of any action taken on their account or content, the user is left without access to a remedy. This censorship is ongoing as we are receiving daily cases documenting consistent patterns.

Based on our findings, we are concerned about the number of cases in which it appears that erroneous enforcement of Meta's policies disproportionately targets content in support of Palestine, effectively silencing Palestinian journalists, activists, and others, as well as voices raising awareness about human rights abuses in Palestine during a critical time.

In 2021, Human Rights Watch published a report which documented Facebook's censorship of the discussion of rights issues pertaining to Israel and Palestine, and called for the company to take up the Facebook Oversight Board's recommendation to commission an independent investigation. Facebook then commissioned an investigation into its human rights' due diligence in relation to the May 2021 events in Israel and Palestine. In September 2022, the Business for Social Responsibility (BSR) published its findings in the report "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021." Among other findings, the report concluded that Meta's actions in May 2021 "appear to have had an adverse human rights impact on the rights of Palestinian users to freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred." The BSR report identified various instances where Meta policy and practice, combined with broader external dynamics, lead to different human rights impacts on Palestinian as well as other Arabic-speaking users.

We will publish a report on our findings in early December. We acknowledge the documents Meta sent us, specifically its September 2023 update on Israel and Palestine Human Rights Due Diligence and its newsroom post on ongoing work in response to the crisis. In order for our report to be as complete as possible, and to provide Meta with an opportunity to comment, we would appreciate your responses to the following questions in writing:

1) How does Meta explain the suppression of content related to Palestine detailed above? What steps, if any, does Meta plan to undertake to investigate whether it is arbitrarily censoring content related to Palestine?
2) What steps, if any, has Meta undertaken to consider and assess the impact of changes in its recommendation algorithms on Instagram and Facebook referenced in its October 13 newsroom post on the freedom of expression of its users? Specifically:

- Before adjusting or introducing temporary measures to take swift action on violating content, did Meta conduct an impact assessment of such measures on the free expression of Palestinian and pro-Palestinian voices on its platforms?
- Can you confirm the Wall Street Journal's reporting that Meta lowered the threshold to hide comments from users in the Occupied Palestinian Territory if its automated detection systems determined there was at least a 25% chance they violated rules?
- Given that such measures are expected to lead to erroneous takedowns of speech that doesn't violate Meta's policies, did Meta determine what error rates it deems to be acceptable?
- Has Meta seen an increase in appeals of its content moderation decisions concerning Palestinian and pro-Palestine content? What percentage of these appeals have been effective?
- Given that these measures are timebound, how will Meta decide when to lift them?

3) Has Meta made temporary changes to its strikes policy in response to the current hostilities? If so, please provide a revised policy, including which policy violations will result in which account restrictions.
4) Has Meta considered notifying users that their content is being demoted or that its distribution is reduced based on policy considerations, in order to provide the user with the opportunity to appeal these decisions?
5) What steps, if any, has Meta undertaken to determine the percentage of content flagged by automation for violating Meta's policies, including on dangerous organizations and individuals, that has been found to be in error? In these cases, what additional steps has Meta taken to provide users whose content was wrongfully removed with a remedy?
6) Meta has referred to violations of its "most serious policies" as leading to restricting a violating user's ability to use Live on Instagram and Facebook for prolonged periods from their first offense, which Meta underpinned with the example of "someone who shares a link to a statement from a terrorist group with no context." However, Human Rights Watch has documented dozens of cases where Meta imposed such restrictions when users merely shared posts documenting hostilities in Gaza, sharing on-the-ground developments, or expressing solidarity with Palestinians. In these circumstances:

- How does Meta explain the prevalence of such disproportionate restrictions?
- Please provide a reasoned justification to what Meta considers to be its "most serious policies." What are the parameters that determine the seriousness of one policy over another?
- Has Meta seen an increase in appeals regarding such restrictions since October 7? If so, what percentage of those appeals has been successful, and what types of remedy are being provided to users whose accounts have been wrongfully restricted from accessing features such as Instagram and Facebook Live?

7) How does Meta explain some users' reports of their inability to access appeal mechanisms on Instagram and Facebook?
8) Has Meta adjusted its internal guidelines on enforcing its policies in response to the October 7 hostilities? If so, could you share them with Human Rights Watch?

9) The Israeli Cyber Unit, based within the State Attorney's Office, flags and submits requests to social media companies to "voluntarily" remove content. Instead of going through the legal process of filing a court order based on Israeli criminal law to take down online content, the Cyber Unit makes appeals directly to platforms, including Meta, based on their own terms of service. Meta committed in 2021 to "disclose the number of formal reports received from government entities (including the Israel State Attorney Office) about content that is not illegal but potentially violates Meta content policies." Can you provide Human Rights Watch with those statistics? If not, why not?

10) What steps, if any, has Meta undertaken or does Meta plan to undertake to institute any firewalls to prevent undue influence of its public policy staff, including former Israeli and other government officials, over content moderation decisions regarding Israel and Palestine?

We request that you provide us with a response to these questions and any other information that you wish to provide **in writing by December 6** so that we may have the opportunity to review and reflect it as appropriate in our reporting. We may post on our website relevant parts of any response received after that date.

Thank you in advance for your engagement. Should you have any clarificatory questions or comments, please feel free to reach out to my colleague Deborah Brown, Senior Researcher on Technology and Human Rights, at [redacted].

Sincerely,

Tirana Hassan
Executive Director
Human Rights Watch

Cc:
Iain Levine, Senior Human Rights Advisor
Miranda Sissons, Director of Human Rights
Alex Warofka, Human Rights Manager and Emerging Risks Lead

AMMAN - AMSTERDAM · BEIRUT · BERLIN · BRUSSELS · CHICAGO - COPENHAGEN · GENEVA · GOMA · JOHANNESBURG · KYIV · KINSHASA · LONDON · LOS ANGELES · MIAMI · NAIROBI · NEW YORK · PARIS · SAN FRANCISCO · SÃO PAULO · SILICON VALLEY · STOCKHOLM · SYDNEY · TOKYO · TORONTO · WASHINGTON · ZÜRICH

55   HUMAN RIGHTS WATCH | DECEMBER 2023

Annex II: Meta's Response Letter to Human Rights Watch

Meta

December 6, 2023

Tirana Hassan
Executive Director
Human Rights Watch
350 Fifth Avenue, 34th floor
New York, NY 10118-3299
United States

Dear Ms Hassan,

Thank you very much for your letter of 15 November 2023. We deeply appreciate the work of Human Rights Watch to document violations and advance human rights, often in the most difficult circumstances.

In the wake of the 7 October terrorist attacks in Israel, Meta took immediate crisis response measures. As we did so, we were guided by core human rights principles, including respect for the right to life and security of the person, the protection of the dignity of victims, and the right to non-discrimination – as well as balancing those with the right to freedom of expression. We looked to the UN Guiding Principles on Business and Human Rights to prioritize and mitigate the most salient human rights risks: in this case, that people may use Meta platforms to further inflame an already violent conflict. We also looked to international humanitarian law (IHL) as an important source of reference for assessing online conduct.

As conflict-related content exponentially surged on our platforms, we implemented a number of temporary measures across both Arabic and Hebrew markets, seeking equitable outcomes in limiting prevalence of violating content on our platforms. We note we've sought to make these measures time-bound, proportionate, and consistent with established human rights guidance on permissible limitations for freedom of expression.

We publicly shared details in our newsroom on October 13 and updated on October 18 and December 5 in English, Arabic, and Hebrew, and have attached a copy to this letter.

Our response has benefited significantly from our prior crisis experience, as well as from the human rights due diligence we conducted and disclosed in 2022, and are continuing to implement (see our most recent update here). For example, we were able to route Arabic

content more appropriately across our systems to enable greater precision, and also to use our [Crisis Policy Protocol](#) to guide our actions.

Obviously, in exceptional and fast-moving situations like this, our response can never be perfect, lines are difficult to draw, and people or systems can and will make mistakes. We know that our users in both the Arabic and Hebrew markets have felt deeply impacted by our decisions. In this respect, we note that none of your questions appear to relate to Hebrew-related markets, measures, or user feedback.

As the conflict evolves, we'll no doubt face new challenges and will continue to learn from civil society feedback and insights from this current period. And we have actively sought out and have heard strong viewpoints from civil society and UN actors in the region and beyond.

We're providing further detail below the signature line – including on our continuing efforts to improve our response. Before doing so we'd like to note: a number of your questions ask us for our opinion on content and/or accounts we have not seen. It is important that we see the content and conduct an assessment based upon all of the available context to reach a determination on whether it violates our Community Standards or Community Guidelines.

We are more than happy to review content samples with you, or to escalate them through our systems, as we are doing daily for many individuals and groups across all relevant markets and languages. Indeed, there are significant benefits in conducting such escalations directly and in real time, both to remedy errors and to share context on specific decisions or policies.

We plan to report back in future on our continuing efforts to address the recommendations made by Business for Social Responsibility in the Israel Palestine human rights due diligence, including any learnings from our current efforts.

Yours sincerely,

Miranda Sissons
Director of Human Rights Policy

Attached:
Response Details
[Newsroom Post](#) (English. [Arabic here](#) and [Hebrew here](#)).

Response Details

Our overall approach

During the current conflict, there has been a surge in related content on our platforms, including hate speech, violence and incitement, dangerous organizations and individuals, and violent and graphic content. While our platforms are designed to support voice, we must also seek to ensure the safety and well-being of our community – and also respond to adversarial coordinated behaviors.

The balance between voice and safety is often not easy to strike in peaceful contexts. In conflict situations – and especially conflict situations involving sanctioned entities, such as Hamas – it is much more difficult.

Content volumes have risen sharply, as has the amount of content people are reporting to us for review. These significantly higher volumes mean we're aware that content that doesn't violate our policies could be removed in error, and that even a small margin of error at scale may result in thousands of negative impacts. To mitigate these risks, we've introduced an important fairness mitigation: for some policy areas, we are temporarily removing violating content without enforcing user strikes, meaning these content removals won't cause accounts to be disabled.

For example, certain posts that violate our Violent and Graphic Content policies may be removed without a strike to ensure that we are not overly penalizing users who are trying to share graphic content and raise awareness about the casualties of war.

We also continue to provide tools for users to appeal our decisions if they think we made a mistake. We have provided more details regarding these steps in our newsroom post.

Hostage video content

In fast-moving crisis or conflict situations, we may need to adapt our guidelines as a situation evolves. This enables our content reviewers to address emerging issues, and also ensure consistency in the application of our Community Standards and Community Guidelines.

In the context of the Israel-Gaza conflict, we initially expanded our Violence and Incitement policy to remove content that depicted identifiable hostages being kidnapped or held in captivity, even when it was done to condemn or raise awareness of their situation. We did so to protect the dignity of the victims in line with international humanitarian law standards (rule against public curiosity in the Geneva Conventions), and also to ensure that Hamas propaganda was not appearing on our platforms in line with our Dangerous Organizations and Individuals policy.

This approach was based on feedback from international humanitarian law experts and civil

3

society in Israel that have raised concerns about the use of footage produced by Hamas showing hostages in captivity.

As the war continues, we have seen the reasons for which people may be sharing hostage related content evolve. In particular, people began sharing hostage content to rebut emerging narratives that the October 7 kidnappings did not actually take place.

We adopted a more nuanced approach accordingly: content showing moment-of-kidnapping footage with speech that condemns the act, or which includes information for awareness-raising purposes is now allowed, with the addition of a warning screen to inform users that it may be disturbing (for example, this could include activists raising awareness of instances of violence or journalists reporting from conflict zones). Under our privacy policy, we may also remove hostage release imagery of minors (under 13). We continue to remove any Hamas-produced footage, unless it is shared in a condemning or news reporting context.

As always, teams are considering context around this imagery, and newsworthy allowances are available where appropriate to balance the public interest against the risk of harm.

You asked if we could share our internal guidelines with you. We want to be clear that Meta does not share internal guidance with external parties unless legally bound to do so.

Appeals

Where a decision is eligible for appeal, people are given the option to ask us to take another look after receiving a notification that their content has been removed or covered with a warning. These appeal mechanisms are available on both Facebook and Instagram. Both human review teams and technology play a role in reviewing user reports and appeals, and we aim to prioritize appeals with potentially harmful content first. During busy periods, such as during conflict situations, we may not always be able to review everything based on our review capacity. Appeals for content demotions are currently not available outside of the EU.

In response to your question about how we define our severest policies, our severe policies refer to those that have a heightened risk of imminent physical harm to individual or public safety. These are assessed and frequently updated to reflect external user research, input from industry stakeholders, safety experts, and law enforcement, as well as a reflection of Meta's company values. These policies would typically include posting content involving terrorism, child exploitation, human trafficking, suicide promotion, sexual exploitation, the sale of non-medical drugs or the promotion of dangerous individuals and organizations.

Temporary measures to help people stay safe

During critical moments with elevated risk of violence or other severe human rights risks, we rigorously seek to tailor our approach to keeping people safe while respecting their ability to express themselves. We closely monitor real world events and platform trends and track different metrics: for example, how much violating content is on Facebook or Instagram, or

4

whether we're starting to see new forms of abuse where we need to quickly adjust our response.

As we mentioned in our newsroom post, we decided to adopt a number of temporary measures in order to keep our users safe and mitigate the risks that our platforms could be used to further exacerbate tensions both online and offline.

We do not implement temporary measures lightly — we know that there could be unintended consequences, like inadvertently limiting harmless or even helpful content. That's why we seek to take steps that are time limited and proportionate to the risks as we are aware of them.

We use machine learning classifiers to identify harmful content and automatically action content when we have high confidence it violates our policies. In emergencies, we may lower the confidence level at which we automatically take action, as we did in this conflict, to address a persistent spike in potentially hateful comments across the region. While these measures are disruptive, our goal is to seek to effectively mitigate salient human rights risks, including the right to life, liberty and security of person; the right to non-discrimination; and others. (For more information on Meta's most salient human rights risks, please see our most recent human rights report, pages 14-26).

Once we see spikes in potentially harmful content return to normal levels and determine the risk to safety on or off our platform has subsided, we will turn off the associated temporary measures.

We note that in conflicts, there are also frequent technical and connectivity challenges that impair user experience, but may result from technical limitations; actions of third parties; or impacts on bandwidth and infrastructure. We know you are aware of the Instagram technical problem on October 16 that impacted users' reach equally across the world, and was fixed as quickly as possible.

Government takedown requests

We do not remove content simply because a government entity (or anyone) requests it. When we receive a content takedown request from a government entity, we evaluate the content in the same way we would for any other piece of content. We first review it against the Facebook Community Standards or Instagram Community Guidelines and, if we determine that the content goes against our policies, we take action. When governments report content that does not go against our policies but is alleged to violate local law, we may restrict access to the content in the country where it's alleged to be illegal, following a careful legal review and a human rights assessment.

As we noted in our September 2023 Israel Palestine Human Rights Due Diligence update, we are still in the process of developing consistent and reliable systems for gathering metrics on the number of pieces of content removed under the Community Standards as a result of government requests. The objective is to produce government takedown request metrics in

5

the most efficient manner given ongoing challenges including confidentiality obligations and data logging and taxonomy gaps from internal systems. We continue to evaluate approaches to building the necessary internal data logging infrastructure to enable us to publicly report this information across the diversity of request formats in which we receive it, but we expect it to be a complex, long-term project.

You asked whether we would be instituting any firewalls between our Public Policy staff, "including former Israeli and other government officials" to prevent any undue influence over content moderation decisions. The idea that the background of our team members has an inappropriate influence on the company's content decisions and enforcement is misplaced and inaccurate: employees of different backgrounds from around the world, including Israelis and Palestinians, are represented in teams working across the company - a fact that's been confirmed by the independent human rights due diligence exercise we commissioned from Business for Social Responsibility.

All Meta employees have to respect our Code of Conduct, which defines the expectations for how we act and how we make decisions. The Code includes specific provisions on human rights and on interacting with governments and political entities responsibly. Our Public Policy staff help us better understand the realities of the situation on the ground for the countries they cover. They regularly engage with government officials, lawmakers, regulators and civil society stakeholders to foster greater understanding of Meta's products, policies and positions across a wide range of issues. The Content Policy team is responsible for writing our Community Standards and deciding how these should be applied in crisis situations. The team is made up of subject matter experts who consult with a wide range of internal cross-functional partners, external experts where appropriate and utilize experience from previous events to understand the specific context to reach these decisions.

Our continuing efforts to improve our response

Since the onset of the current conflict, Meta has been conducting weekly investigations to provide directional analysis to assess the performance of our content moderation systems. The sampling methodology has been designed to assess whether there are issues of over- and under-enforcement on the conflict-related content based on the internal data sets and the submissions by external partners. These included content in Arabic, Hebrew, and English. A dedicated team identifies core issues causing misalignment with what is an expected performance of Meta systems and conducts root cause analysis to identify the solutions to fix them.

The assessment about Meta's enforcement accuracy on entities related to the Israel-Gaza conflict has been conducted in two parts, with slight changes to the company's sampling methodology to capture perceived instances of under-enforcement as well as over-enforcement. Throughout the process, sampled content was derived from review of viral content in relevant policy areas based on Viewer per Views (VPVs), a review of a subset of actioned viral content in those areas, content escalated through user reporting and captured

6

by Quality Assurance, and a samples of content escalated to Meta by external parties that were perceived to be over enforcement.

When internal teams come across an enforcement error, we work to either restore (or remove) the content based on additional review. This work is ongoing.

We plan to report back on our continuing efforts to address the recommendations made by Business for Social Responsibility in the Israel Palestine human rights due diligence, including any learnings from our current monitoring of the conflict.

RESPONSE DETAIL ENDS
REMAINDER OF PAGE INTENTIONALLY LEFT BLANK

7

Meta

Meta's Ongoing Efforts Regarding the Israel-Hamas War

October 13, 2023
Hebrew translation, *Arabic translation*

Update on December 5, 2023 at 9:00PM PT:

At the beginning of the war, we designated the October 7 attack by Hamas as a Terrorist Attack under our Dangerous Organization and Individuals policy. Consistent with that designation, we removed all content showing identifiable victims at the moment of the attack.

Following that, people began sharing this type of footage in order to raise awareness and condemn the attacks. Meta's goal is to allow people to express themselves while still removing harmful content. In turn, we began allowing people to post this type of footage within that context only, with the addition of a warning screen to inform users that it may be disturbing. If the user's intent in sharing the content is unclear, we err on the side of safety and remove it.

Under our Dangerous Organizations and Individuals policy, we continue to remove any imagery that is produced by a Dangerous Organization or Individual, unless it is clear that the user is sharing it in a news reporting or condemnation context, and no minors under thirteen years old are depicted.

Update on October 18, 2023 at 3:00AM PT:

After the terrorist attack by Hamas against Israel last week, and Israel's response in Gaza, our teams introduced a series of measures to address the spike in harmful and potentially harmful content spreading on our platforms. Our policies are designed to keep people safe on our apps while giving everyone a voice. We apply these policies equally around the world and there is no truth to the suggestion that we are deliberately suppressing voice. However, content containing praise for Hamas, which is designated by Meta as a Dangerous Organization, or violent and graphic content, for example, is not allowed on our platforms. We can make errors and that is why we offer an appeals process for people to tell us when they think we have made the wrong decision, so we can look into it.

Some additional updates and steps we are taking as this situation continues to unfold:

Fixing Bugs: We identified and fixed some bugs this past week.

- One impacting all Stories that re-shared Reels and Feed posts on Instagram, meaning they weren't showing up properly in people's Stories, leading to significantly reduced reach. This bug affected accounts equally around the globe – not only people trying to post about what's happening in Israel and Gaza – and it had nothing to do with the subject matter of the content. We fixed this bug as quickly as possible.

- Another prevented people from going Live on Facebook for a short time. This was also a global issue that was fixed within a few hours. We understand people rely on these tools and we're sorry to anyone who felt the impact of these issues.

Comment and Profile Settings:

- As a temporary measure to protect people in the region from potentially unwelcome or unwanted comments, we have:

  - Changed the default setting for who can comment on newly created public Facebook posts of people in the region to Friends and/or established followers only. Users globally can choose to use this setting and opt in or out at any time, and we are notifying people in the region with specific instructions on how to change this setting.

  - We've made it easier for people to bulk delete comments on their posts.

  - Disabled the feature that normally displays the first one or two comments under posts in Feed.

- We recently rolled out the Lock Your Profile tool in the region that allows people to lock their Facebook profile in one step. When someone's profile is locked, people who aren't their friends can't download, enlarge or share their profile photo, nor can they see posts or other photos on someone's profile, regardless of when they may have posted it.

- Fundraising on Facebook and Instagram: Since October 7, people have raised more than $11.5 million for nonprofits on Facebook and Instagram to help with relief efforts in Israel and Palestine. This includes over 340,000 donations to 262 charities – providing disaster relief, ambulance and blood services, medical care and more.

Originally published on October 13, 2023 at 1:00AM PT:

Like many, we were shocked and horrified by the brutal terrorist attacks by Hamas, and our thoughts go out to civilians who are suffering in Israel and Gaza as the violence continues to unfold.

Since the terrorist attacks by Hamas on Israel on Saturday, and Israel's response in Gaza, expert teams from across our company have been working around the clock to monitor our platforms, while protecting people's ability to use our apps to shed light on important developments happening on the ground. The following are some of the specific steps we have taken:

Taking Action on Violating Content

- We quickly established a special operations center staffed with experts, including fluent Hebrew and Arabic speakers, to closely monitor and respond to this rapidly evolving situation in real time. This allows us to remove content that violates our Community Standards or Community Guidelines faster, and serves as another line of defense against misinformation.

- We continue to enforce our policies around Dangerous Organizations and Individuals, Violent and Graphic Content, Hate Speech, Violence and Incitement, Bullying and Harassment, and Coordinating Harm.
  - In the three days following October 7, we removed or marked as disturbing more than 795,000 pieces of content for violating these policies in Hebrew and Arabic.
  - As compared to the two months prior, in the three days following October 7, we have removed seven times as many pieces of content on a daily basis for violating our Dangerous Organizations and Individuals policy in Hebrew and Arabic alone.

- Hamas is designated by the US government as both a Foreign Terrorist Organisation and Specially Designated Global Terrorists. It is also designated under Meta's Dangerous Organizations and Individuals policy. This means Hamas is banned from our platforms, and we remove praise and substantive support of them when we become aware of it, while continuing to allow social and political discourse — such as news reporting, human rights related issues, or academic, neutral and condemning discussion.

Annex III: Human Rights Watch Letter to the Israeli Cyber Unit at the Office of the State Attorney

350 Fifth Avenue, 34th Floor
New York, NY 10118-3299
Tel: +1-212-290-4700
Fax: +1-212-736-1300; 917-591-3452

November 22, 2023

Dr. Haim Vismonsky
Director of the Cyber Unit
Office of the State Attorney
Ministry of Justice
29 Salah A-Din, Jerusalem
9711052, Israel

Dear Dr. Haim Vismonsky,

We are writing to you on behalf of Human Rights Watch, an independent, nongovernmental organization that monitors and reports on human rights in over 100 countries around the world, including Israel and Palestine, where we have worked for more than three decades.

Human Rights Watch has been documenting online censorship related to Israel and Palestine since October 7. In October, Human Rights Watch published a call for evidence of online censorship and suppression of content related to Israel and Palestine on social media since October 7. The call was posted from the Human Rights Watch main accounts in English, Arabic, and Hebrew, on multiple social media platforms, including Instagram, Facebook, X, and TikTok.

Since then, we have gathered hundreds of cases which detail various forms of censorship impacting posts and accounts documenting, condemning, and raising awareness about the recent and ongoing hostilities in Israel and Palestine, which involve serious human rights abuses.

We note that the Israeli Cyber Unit flags and submits requests to social media companies to "voluntarily" remove content, by making direct appeals to platforms based on their own terms of service, rather than filing a court order based on Israeli criminal law.

In order for the report to reflect information from and the perspective of the Cyber Unit, we would appreciate your responses to the following queries in writing:

1. Provide the following details on the Cyber Unit's requests to social media companies since October 7:
   - Number of requests per platform (broken down by number of posts or accounts)
   - Compliance rate per platform
   - The relevant company policy the post or account allegedly violated
   - Compliance rate per policy

2. Please provide the same details on the Cyber Unit's requests to social media companies for the months of October and November 2022.

3. What standards or guidelines does the Cyber Unit follow in making voluntary requests to social media companies? Who determines that a request should be made? How are these determinations made?

We request that you provide us with a response to these questions and any other information that you wish to provide **in writing by December 6, 2023,** so that we may have the opportunity to review and reflect it as appropriate in our reporting. We may post on our website relevant parts of any response received after that date.

Thank you in advance for your engagement. Should you have any clarifying questions or comments, please feel free to reach out to me at [redacted].

Sincerely,

Deborah Brown
Acting Associate Director, Technology and Human Rights Division
Human Rights Watch

Illustration © 2023 Human Rights Watch

hrw.org